Cirad

Université Montpellier 2

Thèse de doctorat de l'Université de Montpellier 2

École Doctorale : Science des Procédés, Science des Aliments

présentée par

Matthieu Papillaud

Contribution au développement de méthodes d'étalonnage à la spectroscopie Terahertz sur des produits biologiques

Soutenance prévue le 15 décembre 2011 devant le jury composé de :

M. Jean-Michel ROGER : .....Directeur de thèse
Mme Csilla GERGELY : .....Co-directrice de thèse
M. Douglas RUTLEDGE : .....Rapporteur
M. Patrick MOUNAIX : .....Rapporteur
M. Jean-Louis COUTAZ : .....Président du jury
M. Max REYNES : .....Examinateur

Résumé :

Ce manuscrit traite de la contribution aux méthodes d'étalonnage à la spectroscopie Terahertz (THz) et de la caractérisation et de la quantification de produits pulvérulents par spectroscopie THz. Le sujet a été orienté afin de fournir les études préliminaires nécessaires à une thématique visant la détection de contaminants sur les aliments, à savoir la caractérisation métrologique de l'appareil (études de répétabilité, sensibilité...) ainsi que la faisabilité de la quantification de produits et l'application de méthodes chimiométriques lors du prétraitement des spectres. La thèse est ordonnée autour de trois publications. La première publication consiste en une revue de littérature servant à faire le point sur les applications concrètes existant en spectroscopie THz. La seconde porte sur la première partie de notre travail, à savoir la caractérisation métrologique du spectromètre THz sur lequel nous avons effectué nos mesures. Enfin, la troisième porte sur l'aspect quantification de la spectroscopie THz et la possibilité d'appliquer les mêmes principes et techniques chimiométriques qu'en spectroscopie infrarouge.

Mots-clés : spectroscopie Terahertz (THz), chimiométrie, évaluation non destructive, métrologie, quantification, étalonnage.

Title:

Contribution to the development of calibration methods to the Terahertz spectroscopy on biological products

Summary:

This manuscript concerns the contribution to the development of calibration methods to Terahertz (THz) spectroscopy, the characterization and the quantification of powder products by THz spectroscopy. The subject has been aimed to give preliminary analysis to a wider thematic of contaminants detection on aliments, which implies the metrological characterization of the device (repeatability, sensitivity...) and the quantification feasibility of these products and the application of chemometrics methods for spectral pretreatment. The thesis is organized around three publications. The first publication is a literature review, which aims to list but a few of the concrete applications of THz spectroscopy. The second one concerns the metrological characterization of the THz spectrometer we worked on. Lastly, the third one deals with the quantification aspect of THz spectroscopy and the possibility of using the same principles and chemometrics techniques that are used in infrared spectroscopy.

Keywords: Terahertz (THz) spectroscopy, chemometrics, non-destructive analysis, metrology, quantification, calibration.

Table des matières

Synopsis

Le présent manuscrit résume le travail que j'ai effectué au cours de ces trois années de thèse au sein des laboratoires du Cirad, du Laboratoire Charles Coulomb de l'Université Montpellier 2 et du Cemagref, sous la direction de Jean-Michel Roger et de Csilla Gergely. Les thématiques de ces laboratoires s'articulent autour de problématiques chimiques, agroalimentaires et (bio)physiques. Ce sujet de thèse au caractère fortement pluridisciplinaire se situe à la frontière entre ces trois domaines. En effet, l'objectif à l'origine de cette thèse a été de développer et appliquer la spectroscopie Terahertz (THz) pour l'identification et la quantification des résidus de pesticides à la surface des fruits et légumes. Toutefois, ce projet a dû être revu suite aux limitations technologiques des instruments et du matériel à disposition à détecter des pesticides sur un support alimentaire. Le sujet a été réorienté afin de fournir les études préliminaires nécessaires à une telle thématique, à savoir la caractérisation métrologique de l'appareil (études de répétabilité, sensibilité...) ainsi que la faisabilité de la quantification de produits et l'application de techniques chimiométriques lors du prétraitement des spectres.

Ce manuscrit est organisé autour de trois publications rédigées et soumises durant la thèse. La première publication consiste en une revue de littérature servant à faire le point sur les applications concrètes existant en spectroscopie THz. La seconde porte sur la première partie de notre travail, à savoir la caractérisation métrologique du spectromètre THz sur lequel nous avons effectué nos mesures. Enfin, la troisième porte sur l'aspect quantification de la spectroscopie THz et s'il est possible d'appliquer les mêmes principes, les mêmes techniques chimiométriques qu'en spectroscopie infrarouge.

Remerciements

Il y a beaucoup de personnes à remercier qui m'ont accompagnées tout au long de cette thèse, aussi je m'excuse par avance auprès de celles que je ne mentionne pas mais que je n'oublie pas pour autant.

A Jean-Michel et Fabrice, mes deux piliers. Sans votre présence, votre soutien, vos encouragements et votre aide, cette thèse ne serait peut-être pas arrivée à son terme. Les mots ne suffisent pas à exprimer ma gratitude. Ça a été un honneur et un plaisir de travailler avec vous.

A Frédéric, Dominique, Philippe, Wojciek, Nina et Oleg, merci pour vos conseils et vos coups de mains, ça a été un plaisir de travailler avec vous mais en premier lieu de faire votre connaissance.

A Emma, Marie, Christian, Amine, les Z'Emilies et les autres thésards, actuels ou passés, pour leur soutien et leur amitié.

A Guy, Csilla et Max, pour avoir lancé le projet.

A Joce et Marie-Pierre, secrétaires de chic et de choc.

A Pierre, Pascaline, Marc, Joël, Gilles et les autres membres du Cirad.

A Fabien, pour ces conseils et ses encouragements malgré son emploi du temps surchargé.

Et surtout à ma famille pour leur indéfectible soutien et pour avoir cru en moi plus souvent que moi-même durant les derniers temps de la thèse.

Ne croyez pas aux miracles. Comptez dessus.

Dixième règle de Finagle

Nomenclature

Les lettres majuscules grasses sont employées pour désigner des matrices, par exemple **X** ; les lettres minuscules grasses désignent des vecteurs colonnes, par exemple $\mathbf{x_j}$ désigne la *j*-ème colonne de **X** ; les vecteurs lignes sont désignés par l'opérateur de transposition, par exemple $\mathbf{x_i}^T$ désigne la *i*-ème ligne de X ; les lettres minuscules non grasses désignent des scalaires, comme des éléments de matrice x_{ij} ou des indices *i*. En cas de besoin, la dimension des matrices peut être indiquée par un double indiçage entre parenthèses : $\mathbf{X}_{(np)}$ indique que la matrice X a *n* lignes et *p* colonnes.

Sauf indication contraire, les notations suivantes sont employées :

X	Une matrice de *n* spectres, par *p* longueurs d'onde	
p	Le nombre de colonnes de **X**	
n	Le nombre de lignes de **X**	
j	Un indice de colonne	
i	Un indice de ligne	
n_{LV}	Nombre de variables (vraies ou latentes) du modèle	
y	Le vecteur des *n* valeurs de référence	
$\hat{\mathbf{y}}$	Le vecteur des *n* valeurs de référence estimées	
b	Le vecteur des *p* coefficients de la régression	
b_0	L'intercept	$\hat{\mathbf{y}} = \mathbf{Xb} + b_0$
$\mathbf{1}_n$	Le vecteur colonne contenant *n* 1	
$\mathbf{I}_p$	La matrice identité de dimension *p*	
$\bar{\boldsymbol{x}}^T$	Le vecteur ligne des moyennes des colonnes de **X**	
$\bar{\boldsymbol{X}}$	La matrice contenant *n* fois $\bar{\boldsymbol{x}}^T$	$\bar{\boldsymbol{X}} = \mathbf{1}_n \ \bar{\boldsymbol{x}}^T$
$\bar{\mathbf{y}}$	La valeur moyenne de **y**	
$\|\mathbf{x}\|$	La norme euclidienne de **x**, i.e. $(\mathbf{x}^T\mathbf{x})^{1/2}$	
SECV	L'erreur standard d'étalonnage en validation croisée	
SEP	L'erreur standard de test	
BS	Le biais de prédiction	
SEP_C	Le SEP corrigé du biais SEP	
Outlier	Individu (i.e. spectre, échantillon) aberrant	

Introduction générale

On appelle le domaine Térahertz (THz) ou lointain infrarouge, la partie du spectre électromagnétique située entre l'infrarouge et les micro-ondes, d'un point de vue fréquentiel, il s'étend de 0,1 THz à une dizaine de THz. En longueur d'onde, le domaine THz va de 30 µm à 3 mm. La zone THz est parfois caractérisée sous l'appellation de "gap THz", du fait du faible développement applicatif de cette zone. On peut d'ores et déjà remarquer que la zone THz se situe à la frontière entre deux domaines de recherche distincts l'un concernant l'électronique (partie ondes hertziennes) et l'autre concernant l'optique comme cela est représenté sur la Figure 1.

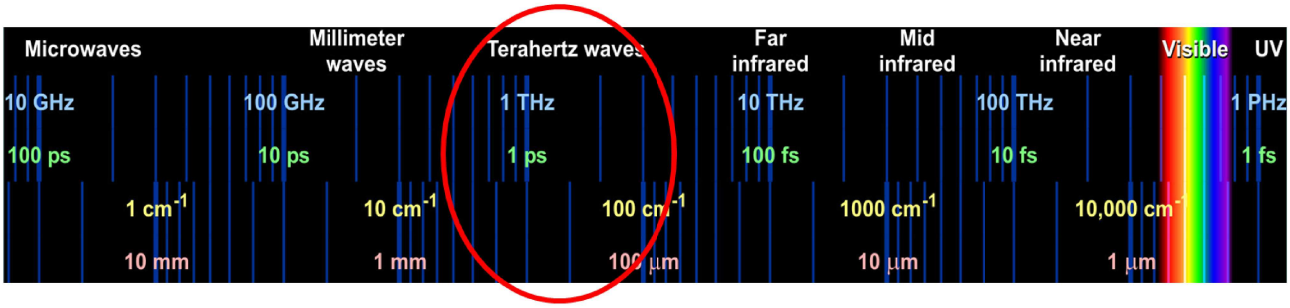

Figure 1: Spectre électromagnétique.

Le spectre électromagnétique est séparé en zones de manière assez artificielle, avec souvent comme référence le domaine visible. Chaque zone, une fois le principe de génération et de détection des photons d'intérêt maîtrisé, connaît un fort développement surtout axé sur la réalisation d'applications grand public. Pour le domaine des rayons X, on peut citer la radiographie médicale X; le domaine UV a donné lieu au développement de la photochimie, de la lithographie et de matériaux filtrant les UV, utilisés dans les lunettes de soleil ou crèmes solaires ; le domaine IR a vu l'explosion de systèmes de commandes sans fil et de la spectroscopie des vibrations moléculaires ; le domaine micro-ondes peut s'enorgueillir du four micro-ondes ou du téléphone portable ; les ondes radio sont à la base des télécommunications ; mais le domaine THz lui ne connaît pas à l'heure actuelle de telles applications essentiellement parce qu'il constitue un domaine de recherche jeune,

en plein développement.

Ce sont essentiellement ces difficultés techniques qui distinguent la gamme térahertz de l'infrarouge lointain, largement couvert par la spectroscopie à transformée de Fourier développée à partir des années 1950 [Griffiths 1986 – Grigoriev 2004]. Le premier article sur le rayonnement térahertz mentionné séparément de l'infrarouge lointain est paru en 1988 [Smith 1988]. Le développement des térahertz à partir du début des années 1980 suit celui des lasers femtosecondes, une composante indispensable des dispositifs térahertz aujourd'hui. En 1996, le terme des rayons T (par analogie aux rayons X) a été proposé dans le cadre des applications à l'imagerie [Mittleman 1996]. Aujourd'hui les ondes térahertz commencent à sortir du monde du laboratoire pour trouver des applications dans la vie de tous les jours, comme les chaînes de contrôle qualité dans les usines ou les portiques de sécurité dans les aéroports.

Spectroscopie aux longueurs d'onde Térahertz

La spectroscopie est l'étude de la structure moléculaire ou atomique et de la composition d'un matériau par la mesure de la radiation électromagnétique qui est absorbée, émise ou diffusée par le matériau sous la forme d'une fonction de la longueur d'onde. La manière dont le rayonnement interagit avec le matériau varie tout au long du spectre électromagnétique, des interactions électroniques aux faibles longueurs d'onde correspondant au domaine des rayons X, jusqu'aux longueurs d'onde plus importantes du domaine micro-ondes. La zone de fréquence THz est située entre les zones micro-onde et infrarouge, ce qui signifie que la spectroscopie THz sonde les liaisons moléculaires, plus spécifiquement les modes vibrationnels et rotationnels. La spectroscopie THz se réfère à des techniques où l'on mesure quelle est la quantité de lumière à une longueur d'onde donnée qui est absorbée, transmise ou réfléchie par un échantillon.

Les liaisons chimiques des molécules ont des fréquences spécifiques auxquelles elles vibrent ; elles correspondent aux niveaux d'énergie de la molécule. Les fréquences vibrationnelles sont reliées à la

force de la liaison et à la masse des atomes à chaque extrémités. La spectroscopie d'absorption est basée sur la comparaison entre l'intensité du rayonnement qui est mesurée avant et après l'interaction avec un échantillon pour fournir à la fois une information qualitative, la composition chimique de l'échantillon, et aussi une information quantitative sur les concentrations relatives des produits absorbants contenus dans l'échantillon. La spectroscopie d'absorption est majoritairement effectuée en transmission, mais elle peut être appliquée à des mesures en réflexion. Elle s'utilise dans le cas d'analyse de produits purs et homogènes ou de mélanges complexes.

Spectromètre Terahertz

Spectroscopie dans le domaine temporel

La spectroscopie dans le domaine temporel (time-domain spectroscopy, TDS, Figure 2) repose sur la mesure du champ électrique térahertz en fonction du temps. Ensuite, une transformée de Fourier permet de passer dans l'espace des fréquences afin d'obtenir le spectre du signal. La réponse du système étant linéaire, le rapport des spectres mesurés avec et sans l'échantillon permet d'éliminer la réponse de l'appareil pour ne garder que celle de l'échantillon.

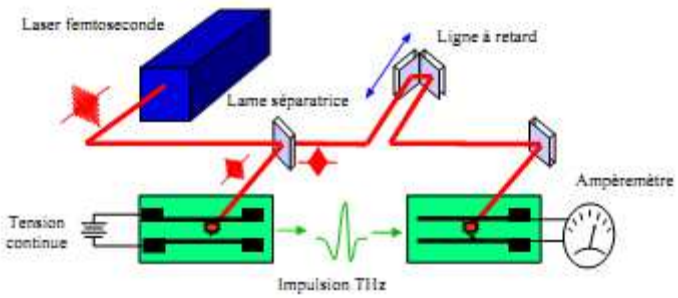

Figure 2: Schéma d'un spectromètre THz dans le domaine temporel.

Génération d'ondes Terahertz

La première exploration quantitative du domaine THz a été menée par Rubens lors de l'étude du rayonnement du corps noir [Rubens 1901]. Toutefois, le domaine THz est resté pendant près d'un siècle privé de sources puissantes. Seuls les corps noirs étaient capables de générer de telles ondes,

et la détection se faisait essentiellement avec les détecteurs pyroélectriques ou avec des bolomètres. On peut également citer, mais de manière plutôt marginale, l'utilisation de lasers moléculaires dans l'infrarouge lointain.

Il a fallu attendre l'arrivée des lasers impulsionnels (lasers produisant des impulsions de lumière dont le champ électrique est confiné dans le temps à des échelles proches de la dizaine de femtosecondes dans le domaine visible par exemple) pour voir un engouement certain de la communauté scientifique pour ce domaine de recherche. En effet la majorité des sources THz impulsionnelles actuelles repose sur l'excitation de différents matériaux par des impulsions laser ultracourtes. Rectification (ou redressement) optique et photoconduction sont les deux approches les plus performantes pour générer des impulsions THz large bande spectrale (allant jusqu'à plusieurs dizaines de THz) [Auston 1975].

Exemples de sources Terahertz

Carcinotron

Les premières sources de rayonnement cohérent sont les tubes micro-ondes, inventés au début du vingtième siècle. Ces tubes, supplantés depuis par les composants à semi-conducteur dans de nombreux domaines de l'électronique, restent aujourd'hui encore indispensables dans le domaine submillimétrique, ainsi que dans les applications nécessitant de fortes puissances (telles que les émetteurs de télévision ou les radars). Ainsi, la seule source commerciale accordable oscillant au dessus de 1 THz est le carcinotron (ou Back Wave Oscillator BWO), qui a été développé dans les années 1970 en France et en Russie. La carcinotron est donc un générateur d'ondes continues qui repose sur le principe suivant : un faisceau monocinétique d'électrons circule parallèlement à une grille. La modulation du potentiel de la grille crée un champ électrique variable, dont la fréquence est liée au pas de la grille. Cette modulation spatiale du champ de la grille module alors l'accélération des électrons qui se mettent alors à rayonner. Le niveau de puissance est compris entre 1 W aux basses fréquences et quelques milliwatts au-delà de 1 THz. Ce type de source est surtout

un instrument de laboratoire dont le coût élevé (40000 euros) et la durée de vie limitée (quelques centaines d'heures) sont rédhibitoires pour son développement. Sa fabrication repose en effet sur de la mécanique de précision devenant micrométrique aux fréquences THz. Le développement de la microtechnologie a toutefois relancé l'intérêt pour les tubes à vide avec des propositions intéressantes visant par exemple la microfabrication de klystrons (tube électronique amplificateur).

Laser à cascade quantique

Le laser à cascade quantique (QCL) est une source laser fonctionnant dans l'infrarouge (IR) ou le lointain infrarouge (FIR ou domaine THz), source basée sur la technologie des jonctions semi-conductrices mais qui diffère notablement des diodes laser. Plus précisément, le développement de l'ingénierie de couches semiconductrices ultrafines (0.5 - 100 nm) a permis de mettre en évidence et d'observer des phénomènes quantiques dans ces structures, phénomènes observés habituellement à des échelles atomiques. Précisément, le laser à cascade quantique est basé sur le confinement quantique et l'effet tunnel.

Le laser à cascade quantique est un excellent exemple de l'utilisation de l'ingénierie quantique pour concevoir des sources électromagnétiques efficaces dans l'IR. Depuis sa première réalisation [Faist 1994], le laser à cascade quantique s'est développé pour fonctionner à plusieurs fréquences, en exploitant des "design" différents, mais utilisant toujours le même matériau : (InP/GaInAs/AlInAs) [Faist 1996-1998]. Ce n'est que récemment que des hétérostructures à base de GaAs/AlGaAs ont été implémentées pour en faire des lasers à cascade quantique [Sirtori 1998]. Le fonctionnement du QCL diffère de celui d'un laser classique en le sens que la cavité résonante optique est remplacée par un effet de cascade quantique assuré par une succession de transitions tunnel entre puits quantiques.

Détection d'ondes THz

En optique, deux grands types de détecteurs sont distingués : les détecteurs quantiques ou optiques

et les détecteurs thermiques. Les premiers sont basés sur l'effet photoélectrique et nécessitent une énergie de photon capable d'engendrer des transitions électroniques. Les détecteurs quantiques transforment à partir d'un seuil l'énergie du photon en énergie électrique. Les seconds transforment l'énergie du photon en chaleur, ils fonctionnent quelle que soit la longueur d'onde du photon mais sont moins sensibles notamment à cause du bruit thermique.

Dans le domaine THz, l'énergie d'un photon est trop faible pour permettre un effet photoélectrique (sauf dans le cas très particulier de semi-conducteurs dopés, on parle alors d'effet photoélectrique interne), aussi il est nécessaire d'utiliser des détecteurs thermiques. Nous n'aborderons ici que le cas du bolomètre, appartenant à la famille des détecteurs thermiques, qui est le détecteur qui a été utilisé durant la thèse.

Figure 3: Bolomètre

Le bolomètre est un détecteur thermique qui fut inventé en 1881 par S. P. Langley [Langley 1881]. Il permet de mesurer la puissance d'une radiation électromagnétique incidente. Le bolomètre utilisé ici est un bolomètre au silicium, il est refroidi à l'hélium liquide, et donc enfermé dans un cryostat. La Figure 3 montre ce cryostat. On aperçoit aussi la fenêtre optique du bolomètre constitué d'une plaque de polyéthylène blanc. Derrière cette plaque, deux filtres additionnels peuvent être placés alternativement, l'un est constitué de polyéthylène blanc dont une des faces est recouverte d'une

couche de 48 µm de diamant (filtre 1) et l'autre de quartz (filtre 2). Cela définit alors deux plages de fonctionnement du bolomètre : le filtre 1 a une plage spectrale comprise entre 0 et 100 cm^{-1} (entre 100 µm et plusieurs mm) et le filtre 2 a une plage spectrale comprise entre 100 et 700 cm^{-1} (entre 14 µm et 100 µm, soit en fréquence entre 3 THz et 21,5 THz). Le bolomètre est relié à un premier étage d'amplification de courant fort gain/faible bruit refroidi lui aussi à la température de l'hélium liquide et à un second étage d'amplification de puissance dont le gain variable est fixé à 200 ou 1000 par l'utilisateur.

Le bolomètre est un détecteur sensible sur une très grande plage de longueur d'onde du rayonnement incident ; un refroidissement à l'hélium liquide à une température de 4 K est nécessaire. Deux autres caractéristiques importantes du bolomètre sont à mettre en avant, sa sensibilité et son faible bruit donné par l'expression signal équivalent au bruit (ou Noise Equivalent Power en anglais, noté NEP) équivalent à 10^{-12} $W.Hz^{-0.5}$. La fréquence d'échantillonage (sampling frequency) d'un bolomètre est de l'ordre de 10 Hz. Tout ceci permet à l'appareil de fonctionner sur une gamme de fréquences inférieure à 30 THz. Le bolomètre est un détecteur quadratique, c'est-à-dire qu'il s'agit d'un détecteur qui délivre un signal proportionnel au carré de l'onde incidente captée. Les détecteurs de type quadratiques sont très amplement utilisés, fonctionnant via la détection et / ou le comptage des photons [Lena 1988].

La chimiométrie

La chimiométrie a été globalement définie par Svante Wold, l'un des pionniers du genre, comme la façon de « *tirer des informations chimiquement pertinentes à partir de données chimiques mesurées, comment représenter et afficher cette information, et comment retrouver cette information dans les données*. » [Wold 1995]

Le but de la chimiométrie peut donc être résumé comme l'analyse de données expérimentales, la visualisation des données et de l'analyse (par exemple des spectres), et la création de plans

d'expériences de façon à maximiser l'information contenue dans les données. La complexité de chacune de ces tâches augmente au fur et à mesure que les données fournies par l'expérience augmentent.

Plus récemment, l'usage du mot « chimiométrie » fait référence à l'utilisation de méthodes de calcul d'algèbre linéaire pour effectuer des mesures d'ordre qualitative ou quantitative de données chimiques, notamment des spectres. La chimiométrie propose aux spectroscopistes de nouvelles méthodes en vue de résoudre le problème de l'étalonnage lors de l'analyse des données. La clé de la compréhension de la chimiométrie n'est pas nécessairement de comprendre les mathématiques impliquées derrière chaque méthode, mais de savoir quel modèle appliquer à un problème donné et comment l'utiliser correctement.

Les applications qualitatives vont s'attarder sur la compréhension d'un modèle, l'identification de molécules, dans le but de comprendre les relations qui existent au sein d'un même système. Les applications quantitatives vont tendre à modeler des propriétés de systèmes chimiques dans le but de prédire les propriétés de nouveaux systèmes testés, introduits dans le modèle. Toutes les applications passent par la constitution de bases de données qui peuvent être de grande taille et très complexes, impliquant plusieurs centaines de variables.

Les domaines de la chimie, de la biologie ou de la physique font appel à la chimiométrie pour traiter leurs données selon leur travail. La spectroscopie vibrationnelle, impliquant plusieurs centaines de variables par spectre (i. e. chaque nombre d'onde) est un domaine de prédilection pour l'utilisation de la chimiométrie. Les informations des produits telles que les concentrations des différents constituants des échantillons analysés peuvent être recoupées avec la forme et l'allure des spectres.

Le but de la démarche est de construire un modèle empirique qui relie un type de données expérimentales (input, comme des spectres), communément désignées matrice **X**, à une type de réponses (output, comme des concentrations), désignées matrice **Y**. Ceci est obtenu en utilisant un set de données, appelé set de calibration, pour lequel les valeurs des données et des réponses sont

connues. Le modèle peut alors être utilisé pour estimer les réponses de **Y** pour de nouvelles données introduites dans **X**. Pour obtenir une estimation de l'incertitude des nouvelles prédictions, le modèle est validé en utilisant un set de données supplémentaire (appelé set de test) avec des matrices **X** et **Y** connues.

Démarche scientifique

La thèse a été réalisée avec le but d'utiliser la spectroscopie Terahertz pour la caractérisation non-invasive de produits agro-alimentaires. Nous nous sommes rendus compte qu'un tel objectif est très ambitieux et présente plusieurs verrous technologiques. Les fruits et légumes sont des systèmes biologiques très complexes non seulement en termes de composition chimique, mais surtout en texture, en structure physique, ce qui peut provoquer une forte diffusion des rayonnements incidents. La présence d'eau est également problématique car l'eau liquide absorbe très fortement le rayonnement THz.

Une étude bibliographique a été menée en premier lieu afin de déterminer quelles étaient les avancées, les applications concrètes effectuées grâce à la spectroscopie THz et nous renseigner sur les possibilités et les limitations de la technique. Il est apparu que la majorité des travaux publiés portaient sur l'étude de produits simples. En chimie, la caractérisation de produits purs prédomine actuellement : la constitution de bases de données spectrales est une étape indispensable pour de futures études. En biologie, l'utilisation de la spectroscopie THz est également orientée vers la caractérisation de biomolécules (ADN, peptides, etc.), soit des produits et des milieux plus complexes. Après ce travail, il fut évident que la détection de produits sur et dans les fruits devait être précédée par la détection, la caractérisation, voire la quantification des constituants d'un aliment.

Or un milieu complexe et riche en eau comme peut l'être une pomme nécessite d'être décomposé dans le but de pouvoir identifier les signaux des différents constituants. Pour retirer le problème du contenu en eau, nous pouvons déshydrater une pomme par exemple, il en résulte une matrice sèche contenant de la cellulose majoritairement, des sucres cristallisés, de l'amidon... sous forme de poudres de produits. La cellulose étant transparente au rayonnement THz [Taday 2004], la problématique était la suivante : est-il possible de quantifier un mélange de sucres sous forme de

poudre dans une matrice transparente ? Pour répondre à cette question, nous avons étudié la faisabilité de quantifier le glucose et le saccharose en poudre dans une matrice de PE, également transparent au THz.

Deux étapes ont été réalisées : la première consistait à caractériser l'appareil du point de vue de sa sensibilité et de sa répétabilité. Cela permet d'identifier d'éventuelles sources de bruit pouvant être traitées par la suite. Ensuite, comme l'exploitation des spectres passe par la création de modèles de prédiction des différents constituants, nous avons étudié l'apport des techniques chimiométriques afin de déterminer si la spectroscopie THz se prête à la quantification de la même manière que d'autres techniques spectroscopiques bien établies comme l'infrarouge.

Ce mémoire est constitué par trois articles dédiés aux trois grandes étapes de notre recherche.

- Le premier est une revue de littérature portant sur les applications chimiques et biologiques de la spectroscopie THz. Les avantages et les limitations de la technique sont présentés et situent le contexte technologique actuel.

- Le second traite de la caractérisation métrologique du spectromètre THz : sa répétabilité, sa sensibilité... Ces aspects sont abordés sous l'angle d'une analyse univariée et sous celui d'une analyse multivariée. Ce dernier nous sert de point de départ pour le développement des méthodes d'étalonnage en démontrant les avantages des techniques multivariées appliquées sur des spectres.

- Le troisième aborde les différentes méthodes d'étalonnage et prétraitements que nous avons testés dans le cadre de la quantification de mélanges de poudres. Il est possible de prédire les proportions en glucose et en saccharose de mélanges à partir d'échantillons purs.

1ère publication : Applications chimiques et biologiques de la spectroscopie Terahertz

Introduction

L'objectif de cette publication est de recenser les applications concrètes en chimie et en biologie, où la spectroscopie THz a été mise à profit. La variété des phénomènes pouvant être analysés en THz est très vaste et de ce fait présente un grand intérêt en recherche pure mais aussi en recherche appliquée. Toutefois, comme cela l'a été dit dans l'introduction, la région du spectre électromagnétique à laquelle la fréquence THz appartient n'a été étudiée que très récemment. Le « THz gap » n'a pu être comblé que depuis une vingtaine d'années grâce à l'apparition de sources et de détecteurs développés expressément pour cette gamme de fréquences. L'évolution constante de ces matériels par le biais du développement des semiconducteurs a permis l'explosion des recherches sur des applications possibles du THz. Pour preuve, une recherche sur Web of Science utilisant les mots-clés terahertz ou THz retourne moins de 20 articles ayant été publiés en 1990, plus de 350 articles en 2000 et près de 1400 articles en 2010. La génération et la détection d'ondes THz font par ailleurs l'objet d'un très grand nombre d'articles mais qui sont plus orientés sur l'aspect physique de la spectroscopie THz. Nous conseillons le lecteur intéressé par cet aspect de consulter principalement le journal *Optics Letters*.

Présentation de l'article

L'intérêt de faire une telle revue de littérature est lié à la relative nouveauté de la technique THz. Comme il s'agit d'une technique en plein développement, tous les champs de recherche se prêtent à l'exploration scientifique. Nous avons décidé de restreindre nos investigations principalement aux domaines biologiques et chimiques car il s'agit des domaines les plus proches du contexte de la thèse. Le champ d'application ayant été exploré est extrêmement vaste : de

l'identification de drogues à la caractérisation de matériel biologique (ADN, protéines, …) en passant par l'imagerie médicale et dentaire. Cette publication n'essaye pas de lister toutes les applications utilisant la spectroscopie THz, mais est plutôt une tentative de faire découvrir cette technique au plus grand nombre considérant les ouvertures scientifiques qu'elle propose.

Discussion

La spectroscopie THz ouvre de nombreuses possibilités dans le domaine de la chimie analytique. Etant donné que la fréquence correspondant au THz concerne les vibrations intramoléculaires mais surtout les vibrations intermoléculaires, les spectres d'un produit peuvent être assimilés à des « empreintes digitales » comme en spectroscopie infrarouge. Toutefois, à la différence de cette dernière méthode, les vibrations intermoléculaires vont permettre de distinguer deux molécules quasi-similaires sur le plan structurel, comme deux isomères différents d'un carbone asymétrique ainsi que le présente la Figure 4.

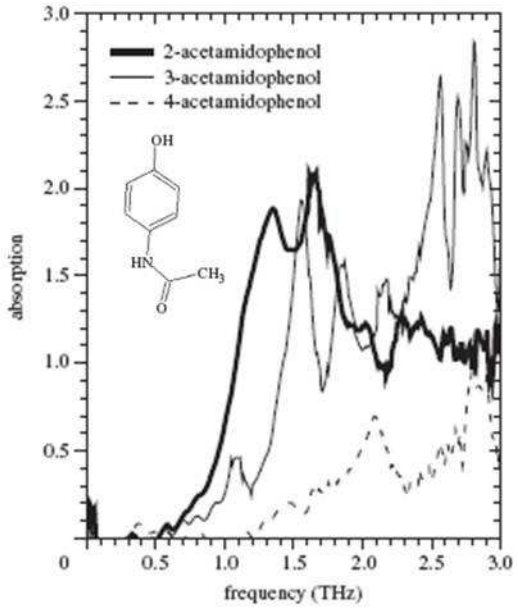

Figure 4: Spectres THz du paracétamol (4-acétamidophénol) et de ses deux isomères. [Taday 2004]

De cette manière, la différenciation des molécules peut être beaucoup plus facilitée par l'utilisation de la spectroscopie THz que par l'utilisation de méthodes spectroscopiques classiques, comme le

proche infrarouge, la RMN ou la spectroscopie Raman. Ainsi, la technique peut être incorporée à des processus de suivi de contrôle afin de surveiller l'apparition ou la disparition de produits de réaction. A ce titre, le THz est en passe d'intégrer les Process Analytical Technology (PAT), soit des méthodes de contrôle en vigueur dans l'industrie pharmaceutique principalement. Par ailleurs, la transparence de nombreux matériaux et objets au rayonnement THz est un atout : la plupart des excipients utilisés en pharmacie n'ont pas de spectre THz, ce qui permet de se concentrer sur les molécules actives. Ainsi, l'enrobage des comprimés peut être analysés via la spectroscopie THz. C'est un point crucial dans l'industrie pharmaceutique puisqu'un mauvais enrobage d'un produit pourrait provoquer la libération anticipée ou au contraire tardive du principe actif, réduisant notablement ses effets. La Figure 5 présente un exemple de spectres ayant servi à contrôler l'enrobage de comprimés, couplée à une analyse par imagerie afin de repérer les éventuels défauts.

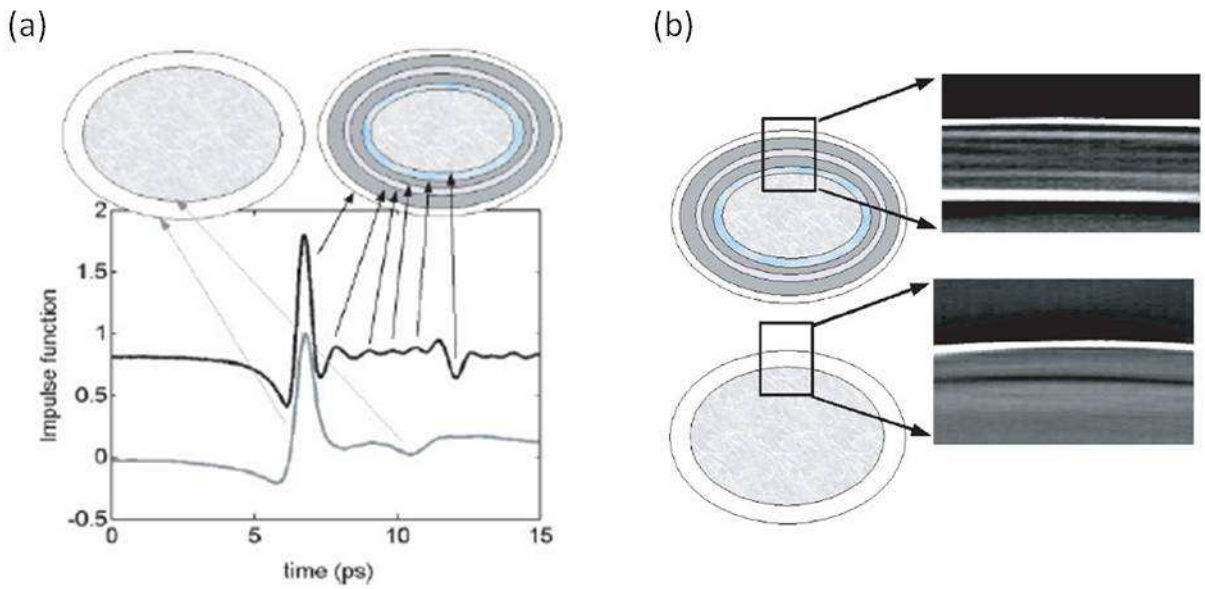

Figure 5: (a) Spectres de deux comprimés comportant plusieurs couches d'enrobage. Chaque bande sur les spectres correspond à une couche traversée par le rayon. (b) Analyse par imagerie de la coupe des comprimés. S'il y a un défaut dans l'enrobage, l'imagerie permet de le situer ; ici on note sur la figure du haut que les couches ne sont pas continues.

Du point de vue biologique, le THz possède les mêmes atouts qu'en chimie, soit une caractérisation inédite du matériel biologique. Le mode de spectroscopie permet d'étudier les molécules en s'affranchissant de contraintes techniques jusqu'ici obligatoires. Par exemple, la spectroscopie THz

permet d'étudier l'état d'hybridation de l'ADN sans avoir recours à un marqueur radioactif qui dénature à terme les échantillons. Dans la partie des applications biologiques de l'article, nous verrons également que le problème lié à la très forte absorption de l'eau, problème qui est également présent en infrarouge, peut être tourné à l'avantage du THz. Par ailleurs, il faut signaler que si l'eau liquide absorbe presque totalement le rayonnement THz, la vapeur d'eau apparaît sur les spectres sous la forme de bandes très fines dont la position est connue. Elles peuvent donc être retirés artificiellement. Les vibrations intermoléculaires dans l'eau solide sont extrêmement atténuées ; il en résulte que la glace est transparente aux rayons THz.

Figure 6: Scan corporel effectué par un appareil développé par la société Qinetiq. Le couteau en céramique dissimulé dans le journal apparaît nettement.

Le domaine de l'imagerie THz est en pleine expansion. Ses applications se trouvent à la fois dans le domaine chimique (contrôle de l'enrobage de comprimés) et dans le domaine biologique (imagerie dentaire, détection de mélanomes), en passant par le domaine de la sécurité. De nombreuses compagnies se sont spécialisées dans la fabrication d'appareillage THz permettant de scanner et imager des objets de diverses tailles. L'application la plus notoire étant les scanners corporels utilisés depuis peu dans les aéroports en vue de détecter rapidement des objets qui n'apparaissent

pas avec les méthodes classiques (détecteur de métaux, rayons X), comme l'illustre la Figure 6. Le rayonnement non nocif et non invasif du THz et le fait que les vêtements et certaines matières plastiques soient transparents au THz font que cette technique se pose comme une alternative viable aux méthodes précédemment citées.

Nous pourrons également noter que le domaine dentaire a été l'un des premiers domaines à utiliser l'imagerie THz dans des applications concrètes. La forte absorption de l'eau est tournée en avantage pour repérer par imagerie les zones cariées (fortement hygroscopiques) ; cela permet une analyse nette des surfaces attaquées par les caries. Il est également possible de séparer la dentine de l'émail, comme le montre la Figure 7.

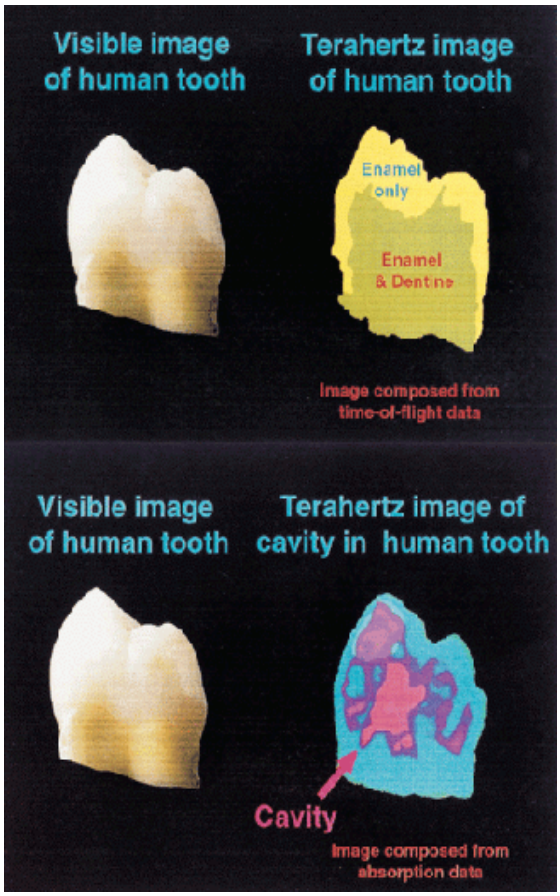

Figure 7: Coupe de dent par imagerie THz. En haut, distinction entre émail et dentine ; en bas, observation d'une cavité à l'intérieur de la dent.

Le présent manuscrit est axé sur la contribution du développement de méthodes d'étalonnages à la

spectroscopie THz, aussi des méthodes chimiométriques vont être abordées dans la suite de l'étude. Toutefois, notre étude bibliographique a révélé que peu de publications ont cherché à appliquer des méthodes chimiométriques au cours de leurs utilisations de la spectroscopie THz. Celles qui le font sont principalement des publications en rapport avec des systèmes simples, chimiques, où la chimiométrie est utilisée à des fins quantitatives [Taday 2004, Strachan 2005, Ge 2006, Wu 2008]. Aucune publication ne s'attarde sur l'exploration des méthodes chimiométriques et son application au THz. Elles se contentent d'appliquer une méthode de régression (principalement la Partial Least Square regression, ou régression PLS) avec peu (dérivée première ou seconde de Savitzky-Golay généralement) ou pas de prétraitement dans le but de tester la faisabilité de la mesure via cette méthode spectroscopique. Il y a donc un champ de recherche à étudier sur les applications multivariées en spectroscopie THz, ce que nous aborderons dès la seconde publication.

Conclusion

Les perspectives en spectroscopie THz sont encore très larges. La jeunesse du domaine fait que presque tout reste encore à explorer. Des projets de recherche ont abouti à la création de bases de données online afin de contribuer à la diffusion et à la compréhension de la technique. Les étapes de recherche suivent les étapes de développement des sources et des détecteurs THz. La spectroscopie THz est en phase de test dans certains milieux industriels dont la médecine, la chimie, la pharmacologie et l'agronomie sont les pionniers. Les avantages de cette technique, ses spécificités, sont autant d'atouts souhaitant être mis à profit par le monde industriel. Si elle n'est pas encore en mesure de supplanter les techniques d'analyse classiques, elle peut contribuer à l'amélioration de certaines analyses (avancement d'une réaction, cristallisation d'une molécule, degré d'hydratation...). Compte-tenu de la constance du développement de matériel THz, la prochaine étape consiste à miniaturiser les systèmes afin de disposer de systèmes réduits pouvant être installés directement sur des lignes de production ou de contrôle.

Publication

as submitted in ***Trends in Analytical Chemistry****.*

CHEMICAL AND BIOLOGICAL APPLICATIONS OF TERAHERTZ SPECTROSCOPY

M. Papillaud*[a], J.M. Roger[d], F. Davrieux[a], C. Gergely[b,c]

[a] Centre International pour la Recherche Agronomique et le Développement (CIRAD), UMR 95 Qualisud, TA B-95/16, 73 rue J. F. Breton 34398 Montpellier Cedex 5, France

[b] Université Montpellier 2, Laboratoire Charles Coulomb UMR 5221, F-34095, Montpellier, France

[c] CNRS, Laboratoire Charles Coulomb UMR 5221, F-34095 Montpellier, France

[d] UMR ITAP, Cemagref, BP 5095, Montpellier Cedex 1, France

*Corresponding author: Matthieu Papillaud, CIRAD UMR Qualisud, TA B-95/16, 73 rue J. F. Breton 34398 Montpellier Cedex 5, France. Phone: +33(0)4 67 61 54 52; Fax: +33(0)4 67 61 44 33. Email: mpapillaud@yahoo.fr

ABSTRACT

Terahertz spectroscopy is a recent and continuously developing technology complementing the range of analytical spectroscopy. Filling a gap in the electromagnetic spectrum, terahertz spectroscopy has issued widespread applications from the identification of drugs to the characterization of biological material through tooth imaging. This paper does not propose to list every application of terahertz, but offers to screen a panel of the main domains using this technique. It is an attempt to raise the awareness among scientists for this technique, especially considering its opportunities in odontology and biological imaging.

KEYWORDS

Terahertz spectroscopy; far infrared spectroscopy; vibrational spectroscopy; analytical chemistry; biology; imaging.

1. INTRODUCTION

Terahertz (THz) spectroscopy might not be a well-know technique, it is nevertheless becoming a more commonly used tool for many applications. Various domains, such as analytical chemistry, pharmaceutics or biomedicine are using THz radiation to perform analysis that other well-established methods, such as infrared (IR) spectroscopy or X-rays, cannot accomplish. Its success is surely related to the fact that THz frequencies lie at the edge of two distinct domains: partly electronics and partly optics. The purpose of this article is to list but several THz applications covering chemical and biological domains. Much information are provided by the frequency range between 0.3 and 3 THz, which correspond to the far infrared vibrational modes, such as rotational, torsional, inter- and intra-molecular vibrational modes, and hydrogen bonding stretches.

The THz range (0.1 - 20 THz approximately) exploration was a late development, because spectroscopists lacked adequate sources for a long time. This ‘THz gap’ was not filled until the 1980s when reliable generators and detectors of THz rays started to be developed. While other techniques gained all attention from scientists because of the accessibility of their frequencies, and thus were the cornerstone of spectral characterization (among many other applications), THz had to wait until the mid-60s to see the apparition of high-power lasers [1] and specific band filters suited to this kind of spectroscopy [2]. It was only during the 80s that the first user facilities were constructed and that THz was used in various fields [3]. The study of electronic materials may have been the widest use of THz technology so far: thanks to the characterization of semi-conductors, conducting polymers or superconductors [4], major breakthroughs have occurred in THz waves generation and detection, turning THz into one of the standard characterization methods of this domain. Among the mentioned advantages are: characterization without electric contact [5] and higher gain in sensitivity [6]. From then, the interest for THz radiations has never ceased

to grow and thanks to the quick evolution of material these past years [7, 8], this technique has become a promising addition to the wide array of analytical modalities. It may not supplant methods that have been used for years and are fully understood such as IR, but it may act as a complement which can provide information otherwise inaccessible. The interest in THz spectroscopy has led to the study of many product categories such as drugs [9, 10], explosives [9, 11-14], sugars [15, 16] or pesticides [17].

In the following of this article, the discussion will be focussed on the THz applications in chemical and biological domains, which have intensely tested its possibilities for detection, characterization and prediction. Medicine, chemistry, pharmaceutics, animal and vegetal biology, and other domains have found numerous applications for the characteristic properties of THz waves. This paper will focus on the recently developed analytical methods in those domains, dividing them into two parts: biochemical and biological systems.

1.1 THz spectroscopic techniques

The first studies were carried out with Fourier Transform Far InfraRed (**FTFIR**) spectroscopy which allowed a good material characterization over a broadband spectrum. However, because of the nature of the sources, the main problem of this technique lies on the low signal to noise ratio at low frequencies which implies a lack of spectral resolution [17]. The measuring system is based on a Michelson spectrometer. A broadband spectrum beam is split in two parts by a half-silvered mirror which produces interferences between two parts of the beam. A movable mirror introduces a phase shift (time delay between the two beams), allowing the light coherence to be measured as a function of the time delay. Thanks to the development of electronic THz wave generators and detectors starting with

the work of Auston *et al.* [18], FTFIR was supplanted during the mid 80s by THz Time-Domain Spectroscopy (**THz-TDS**). Its principle is based on the conversion of a femtosecond laser pulse (mostly generated by a titanium-sapphire laser) into a femtosecond THz pulse, as shown in Figure 8. A single pulse can contain frequency range covering the whole THz range. The laser beam is split in two parts, the pump pulse (or detection pulse) and the gate pulse (used to generate the THz pulse). One part of the beam is sent on an optical delay line to analyse and record the THz pulse as a function of time (stroboscopic effect). The THz pulse passes through the sample and is then focused onto the detector. The Fourier Transform of the delay between THz and detection pulses gives both amplitude and phase spectra over a frequency range related to the pulse duration. The access to the phase is one of the advantages of THz-TDS technique. Another advantage is that the signal to noise ratio is better when using THz-TDS under 3 THz, while FTFIR regains advantage above 5 THz [19].

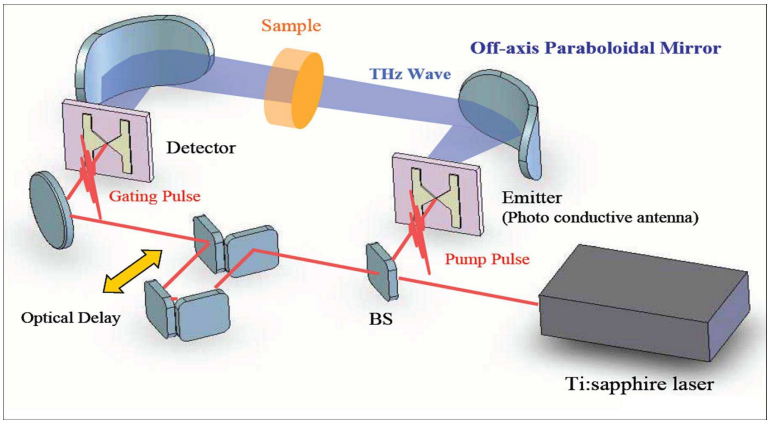

Figure 8: Principle of THz-TDS (courtesy of Riken Sendai)

Terahertz imaging is a different technique which measures the two-dimensional properties of THz spectra: the contrast of the absorption of the sample caused by inter- and intra-molecular interactions. This approach is useful for tomographic studies and for analysing the distribution and concentration of chemicals in targeted materials as Terahertz Pulse Imaging (**TPI**) provide both time of flight analysis for spatial reconstruction

and spectral information. It has to be considered that a 1 THz wave is equivalent to a 300 µm wavelength and to a 4 meV energy. These weak photonic energies explain why THz is considered as a non-invasive technique and has generated much interest. When screening a sample, part of the incident ray will be reflected back towards the detector whenever there is a change in the refracting index of the sample, e. g. crossing different layers, a cavity or different phases, generating a pixel at a time, whose intensity is directly linked to the THz wave phase and amplitude. This method has permitted depth probing and the reconstruction of the internal structure of the sample by combining the reflected pulses to generate a three-dimensional image. Because of its harmless abilities, THz imaging has recently been applied to security devices enabling body imaging: the capacity to detect hidden objects under clothes or in packages at distance has created a whole field of applications of THz technology in this area [20]. Medicine has also seen the potential of THz imaging in many domains. Oncology and odontology have developed applications based on this technique, which will be discussed later in this document.

2. DEVELOPMENT OF THZ APPLICATIONS

Because of THz relative novelty compared to well-established methods, much work has still to be done to master the technique and acquire certain knowledge about spectral interpretation. Several studies have dealt with single products, simply focussing on their acquisition [11, 15, 17] because of specificity of THz spectroscopy: a molecule is entirely excited by THz radiation and not only parts of the molecule given the wavelength as it is the case with NIR spectroscopy for example. As a result, very specific and characteristic spectra are obtained for each molecule. The detection of single products, such as drugs, explosives or various chemicals, and the use of THz technology as an analytical control tool can be grouped into the biochemical applications. These kinds of applications have

become more and more used, especially in the security and pharmaceutical industries which have found a growing interest in THz. On the other hand, biological studies have also used THz to characterize numerous biomolecules (cell constituents, proteins, DNA…) and their activities. Despite the progresses made, applications in more complex biological systems are still very few.

2.1 Simple chemical systems

Analytical chemistry has seen the potential of THz spectroscopy for a long time. Its uses range from gas sensing to molecular characterization and quantification. Pioneers in astronomy have driven technological advances in THz generation and detection. During the late 1960s, astronomers were provided specific spectral frequencies of other galaxies by far-infrared spectroscopy [21]. Following this initial success, the development of this promising technique has resulted in the deep probing of dust clouds and the detection of interstellar molecules: about thirty of these molecules were identified in 1980, now more of 120 are characterized [22]. The latest improvements in THz astronomy have confirmed the presence of water and polycyclic aromatic hydrocarbons in Titan's atmosphere, taking advantage of water strong absorption at THz wavelength [23]. This technology has since been adapted to atmospheric measurements for gas sensing [24].

Scientists have raised the opportunity to use THz for communications: the concerned frequencies are unallocated by Communications Departments and would provided higher wireless data rates. It has been foreseen that THz could provide secure communication links between satellites or even in atmospheric conditions [25, 26]. The material state-of-art is clearly a limiting factor but the first successful data transmission have been realised a few years ago [27].

Drugs and explosives studies exploit a major advantage of THz spectroscopy. There is no background information concerning what conceal the product and illicit or dangerous products can be identified easily. Many of these studies have found an application in the security domain [20]. The characterization of a product by THz spectroscopy has also eased the preparation of samples compared to other techniques. It has been demonstrated that some components, such as polyethylene or cellulose are transparent (or at least semi-transparent) at THz frequencies [28]. Identifications are simplified as there is no superfluous information due to the nature of the matrix that contains the sample, but it has to be mentioned that THz can be sensitive to the preparation method if structural changes happen or impurities are introduced. The chemical sensitivity of the THz technique has contributed to the recognition of different products put in sealed containers without direct contact [29]. This fact has contributed to the success of THz as a new characterization method for security applications, particularly for explosives detection [10-13]. Not only can THz detect and identify explosives, but it also can detect their phase changes, which means it has been able to follow real-time modifications of the molecule [14]. More and more security scanning applications are taking place in airports, the security sector being one of the most receptive to THz technology. Concealed weapons, even if they are made of non-metallic materials such as ceramics, can be quite easily detected through clothes and packages which appear to be transparent at THz wavelengths. For those reasons, THz has also proved to be highly effective at scanning people for hidden objects. However, body imaging is limited by the strong absorption of THz rays by water, so body water prevents in-depth scanning. Small objects concealed in the mouth could not be detected during the scan for example.

Several publications have focused on data acquisition, trying to relate the different peaks

appearing in the spectra to the chemical groups present in the molecule or to predict where the bands should appear by using prediction programs developed for infrared methods. These works have collected valuable data for chemical recognition but have had more diverse results when it came for spectra prediction or interpretation. For example, Li Hor *et al.* have studied trichloroanisole with THz-TDS, acquiring characteristic features but failing to match the predictions made by the computational program they used [30].

A major stake in chemical industry, isomer differentiation can be achieved quite easily and require no further investigation [28]. Therefore, different molecules with even the lightest difference in their molecular structures, which would not be discerned easily with other analysis methods, will get different THz spectra. In this study, Taday has shown that the three molecules analyzed, of very similar structure, present strong features that are not common to every spectrum. The characterization of a molecule does not concern parts or groups that could be found from a spectrum to another; the low vibrational mode concerns the molecule as a whole entity, which explains that the spectra of the three samples have distinctive, characteristic bands. THz spectroscopy is also sensitive to minor changes in molecular conformation. Epimers (diastereomers that differ in configuration of a single asymmetric carbon) have also shown different spectra when characterized by THz [15]. Walther *et al.* have worked in a similar way than chemists, characterizing the non-covalent intermolecular forces in sugars in polycrystalline and amorphous states. The results of this study were satisfying enough, THz enabling the distinction of the different forms quite easily [31]. These studies prove that THz spectroscopy is a very sensitive method for the differentiation of similar products.

Chemical sensitivity has proven to be an asset when samples of unknown composition

have to be identified: the proportion of solids in mixtures can be estimated by THz spectroscopic imaging. Once the spectra of pure products have been acquired, they can be used for imaging recognition. Watanabe *et al.* have set up an experimental design consisting of three pellets made of aspirin, palatinose and a mixture between the two of them, but which measuring order remains unknown [32]. The detecting device was configured to detect one product at a time, which data was previously recorded. Imaging revealed without doubt that the first pellet to be scanned was the aspirin, the second one was the mixture and the last one was composed of palatinose. When the settings were configured to detect riboflavin (quite similar in structure to palatinose), there was no response, meaning that none of the samples contained riboflavin, in acknowledgment of the THz ability for discrimination.

As it has been said before, THz frequencies are concerning intra- and inter-molecular vibrations, as well as hydrogen bonding. Therefore, any difference in materials processing history, i.e. changes in the process that could induce a structural or a conformational change, can be detected by this method and will result in changes among the spectrum.

This makes THz ideal as a Process Analytical Technology as well as IR and NIR in chemical and especially pharmaceutical industries [28]. Many works have transposed IR principles, combining THz and chemometrics pretreatments to evaluate and quantify the proportions between active ingredients and excipients, for example [33]. The levels of polymorphism and crystallinity of drugs have also been studied, quantifying each form during real-time analysis in order to optimize the fabrication process [34]. Reactions are monitored and can be adapted to the situation: diastereomers can be identified and quantified [15], which means that a reaction can be stopped at the moment when all conversions into one form has been done, thus being less time- or product-consuming.

As a spectroscopic technique, THz can be combined to chemometrics to develop prediction and calibration models. The applications cited in the previous paragraph use chemometrics to control the quality of their work. They also confront THz measurements to models created with other techniques such as NIR to see if there are similar behaviours between the methods in order to better understand THz spectra.

Thanks to its non-destructive ability, TPI has been used to investigate the quality of the coating process of pharmaceutical products [35-37]. The deep penetration of THz waves in pellets (as the used excipients are usually transparent to THz) has allowed realising structural mappings of the samples, detecting the different layers that compose them. Whereas NIR is better suited for thin layers and furnish a bigger spatial resolution, THz provides directly the sample thickness, probing deeply the samples as many excipients are transparent or semi-transparent at THz wavelengths [38]. The deep probing, coupled to the THz sensitivity, has made able to control the inside uniformity of pellets in a non destructive manner, opposed to the microscopic analysis that is necessary otherwise. For example, every layer met by the THz radiation would appear as a feature on the spectrum. Thus, it clearly appears when the layers are not homogenous, revealing a default in the coating process [39].

During the formation of polymers, it has been observed that TPI offers a better sensitivity and makes more information available than other techniques [40]. Moreover, the corresponding spectra are totally different depending on the size of the polymers, which make THz perfectly suited to real-time analysis and quantification of a polymeric reaction. The comparison of TPI with NIR spectroscopy for this kind of analysis has shown that TPI is slightly advantageous as it has presented the same kind of results when confronted to prediction models and qualitative analysis, without the need of empirical calibration

development [41].

2.2 Complex biological systems

To our knowledge, biologists mainly use THz spectroscopy as a complement or as a backup for other analytical techniques. A few concrete applications based on THz only have emerged, mainly focussing on imaging.

THz spectroscopy has been used by biologists to contribute to the characterization of biological material. This method is particularly fitted to the study of many cells and cellular organites because their dimensions correspond to the THz wavelength. Many important biological molecules such as DNA [42-44], amino acids [45], proteins [46], polypeptides [47] and sugars [31] have been studied with THz spectroscopy and show strong features. This is due to the fact that numerous biological systems possess vibrational modes for which THz low energy levels are particularly suited.

The desire to observe the conformational structure and the changing states of molecules have led biologists to try a non-destructive and non-invasive method. Among the first biological molecules identified were the nucleobases ACTG and their corresponding nucleosides forming the deoxyribose nucleic acid (DNA). The molecules have been studied independently and in their binding state. The goal of this kind of studies is to lead to a better comprehension of the influence of the environment and of the conformational state of the molecules on DNA. Isolated nucleobases as well as bonded ones have shown very strong and characteristic features. The differences between spectra have been attributed to the inter- and intra-molecular vibrations of hydrogen bonds [43]. Moreover, DNA hybridization has been studied with this technique because of its label-free ability: the

determination of important biological processes such as the changing state of DNA has become a major stake in biology and THz is holding many promises in this domain [48].

Despite the fact that water has shown very strong absorption in the THz region, this inconvenient has been turned into an advantage for tissue identification. Tissues contain different amounts of water, thus having different responses to THz rays. Skin, lean and fat tissues have been measured by THz-TDS, comparing their responses. Experiments have shown that the different tissues from a same animal can be differentiated, whereas the same type of tissue coming from different species has shown very little differences, preventing the identification of animals [49]. It has also been shown that the differentiation between bone and tissue cannot always be easily distinguished. Imaging techniques are then required to fulfil the analysis with enhanced contrasts [50].

THz imaging has also seen many concrete applications for the last years in tissue analysis. TPI has been popularized in oncology, significantly differentiating healthy cells from the cancerous ones. Skin samples have been scanned *in vivo* and *ex vivo*, the differentiation between the cells being resolved thanks to the absorption differences. Indeed, cancerous cells contain more interstitial water than healthy ones, so the absorption of a tumour cell and its refraction index will be higher, easily delimiting the edges of the disease. Despite these encouraging results, this technique cannot be used for prevention yet: the difficulty in characterizing tumour cells among healthy ones is too important in the early stage of the disease. Anyway, thanks to modern instrumentation and the use of chemometrics, information has now been gathered not only at the surface, but also in depth [51]. In the latter case, it has been suggested that some centres as the dermal-epidermal junction, could be source of scattering, dispersing the THz radiation and preventing the collection of all depth information [52]. This factor must be taken in account

for studies involving thick samples containing multiple layers which add to the complexity of a system, especially if the system is hydrated. Recent developments have coupled IR and THz techniques to boost the observed THz signal either by imaging or by monitoring the signal at a point [53]. Gold nanoparticles targeted to the tumours have been used as contrast agent and have been irradiated with an IR laser, leading to a temperature rise in cancerous cells. THz is sensitive to the temperature changes in water so detection of tumour has been made easier. This technique has improved the THz signal up to 30 times higher (in the differential mode) than without the presence of nanoparticles. Thanks to this, Oh *et al.* [53] have obtained resolution gain down to a micron, which could allow deeper probing *in vivo* and detection of cancer at early stage. Davies and Hales have led clinical trials in collaboration with Teraview Ltd in order to develop portable THz spectroscopy system for medical diagnosis, with successful results so far.

Among the other applications of TPI lies dentistry. Researchers have been using THz to scan teeth in order to detect caries. Actual technologies for caries detection use ionising rays such as X-Rays. However, high doses of this radiation are harmful for human beings, so THz spectroscopy stands up as an alternative diagnosis technique. For example, Crawley *et al.* have studied tomography of a decayed tooth [54]. As in oncology, this technique turns the strong absorption of THz rays by water to its advantage. Carious zones are more hydrated than healthy zones. The differences in absorption spectra are combined to tooth scanning to delimit precisely the caries area. Knott *et al.* [55] have worked on the detection of early stage caries in the occlusal enamel layer using TPI. They too noted that healthy enamel had different absorption spectra and contrast when compared to carious enamel and succeeded in the acquisition of an early and safe diagnosis.

3. CONCLUSION

Numerous applications of THz technology exist and only a few domains have been discussed. Thanks to its unique properties, THz is a very promising technique and the constant progress in instrumentation and signal processing will make it a reliable alternative to classical analysis techniques as IR (NIR included), X-rays or Raman spectroscopies. One could make an analogy between IR and THz, comparing the growing interest of the technique when the equipment has started to be user-friendly and more accessible. Researchers have used chemometric algorithms and programs based on IR calculations to predict the position of the bands; not able to correlate the two methods at first, but now succeeding to simulate the response according to multiple factors (water, temperature…) and initiating the identification of the specific vibration modes responsible of it [56].

THz has proven its interest in the characterization of single products, as well as mixtures, but one of the future prospects about its utilization would be the detection and the characterization of relatively simple molecules in the middle of a more complex system.

It should also be mentioned that several projects led to the establishment of online THz database and networks, such as the “Terahertz radiation in Biological Research, Investigation on Diagnostics and study of potential Genotoxic Effects” (http://www.frascati.enea.it/thz-bridge/), the “THz Science and Technology Network” (www.THzNetwork.org), or the database of Riken Sendai (http://www.riken.jp/THzdatabase/) which will contribute to the global spread of THz data and understanding.

REFERENCES

1. A. Crocker, H.A. Gebbie, M.F. Kimmit, L.E.S. Mathias, Nature 201 (1964) 250.

2. R. Ulrich, Infrared Phys. 7 (1967) 37.

3. M.F. Kimmit, Journal of biological physics 29 (2003) 77.

4. M. Hangyo, M. Tani, T. Nagashima, International Journal of Infrared and Millimeter Waves 26 (2005) 1661.

5. P. Planken P, Nature 456 (2008) 454.

6. A.J. Huber, F. Keilmann, J. Wittborn, J. Aizpurua, R. Hillenbrand, Nano Lett. 8 (2008) 3766.

7. B. Ferguson, S. Wang, D. Gray, D. Abbott, X.C. Zhang, Microelectronics Journal 33 (2002) 1043.

8. D. Dragoman, M. Dragoman, Progress in Quantum Electronics 28 (2004) 1.

9. A.G. Davies, A.D. Burnett, W. Fan, E.H. Linfield, J.E. Cunningham, Materials today 11 (2008) 18.

10. J.F. Federici, B. Schulkin, F. Huan, D. Gary, R. Barat, F. Oliveira, D. Zimdars, Semicond. Sci. Technol. 20 (2005) S266.

11. J. Hooper, E. Mitchell, C. Konek, J. Wilkinson, Chem. Phys. Lett. 467 (2009) 309.

12. D.G. Allis, J.A. Zeitler, P.F. Taday, T.M. Korter, Chem. Phys. Lett. 463 (2008) 84.

13. G.F. Liu, X.J. Ma, S.H. Ma, H.W. Zhao, M.W. Ma, M. Ge, W.F. Wang, Chinese Journal of Chemistry 26 (2008) 1257.

14. Y. Hu, P. Huang, L. Guo, X. Wang, C. Zhang, Physics Letters A 359 (2006) 728.

15. M. Ge, H. Zhao, T. Ji, X. Yu, W. Wang, W. Li, W. Science in China: Series B Chemistry 49 (2006) 204.

16. L. Yang, H. Sun, S. Weng, K. Zhao, L. Zhang, G. Zhao, Y. Wang, Y. Xu, X. Lu, C. Zhang, J. Wu, C. Jia'er, Spectrochimica Acta Part A 69 (2008) 160.

17. Y. Zhang, X.H. Peng, Y. Chen, J. Chen, A. Curioni, W. Andreoni, S.K. Nayak, X.C. Zhang, Chem. Phys. Lett. 452 (2008) 59.

18. D.H. Auston, P. LeFur, Appl. Phys. Lett. 28 (1976) 21.

19. P.Y. Han, M. Tani, M. Usami, S. Kono, R. Kersting, X.C. Zhang, J. Appl. Phys. 89 (2001) 2357.

20. NATO Security through Science Series, Terahertz Frequency Detection and Identification of Materials and Objects, second ed., Springer, Netherlands, 2007.

21. F.J. Low, W.H. Tucker, Phys. Rev. Lett. 21 (1968) 1538.

22. M. Rowan-Robinson, Astronomy and Geophysics 47 (2007) 4.31.

23. G. Winnewisser, C. Kramer, Space Science Reviews 90 (1999) 181.

24. D.M. Mittleman, R.H. Jacobsen, R. Neelamani, R.G. Baraniuk, M.C. Nuss, MC. Appl. Phys. B 67 (1998) 379.

25. M.J. Fitch, R. Osiander, John Hopkins Apl Technical Digest 25 (2004) 348.

26. M. Koch, in: R.E. Miles, X.C. Zhang, H. Eisele, A. Krotkus (Eds.), Terahertz Frequency Detection and Identification of Materials and Objects, second ed., Springer, Netherlands, 2007, p. 325.

27. T. Kleine-Ostmann, K. Pierz, G. Hein, P. Dawson, M. Koch, Electron. Lett. 40 (2004) 124.

28. P.F. Taday, R. Soc. Lond. A 362 (2004) 351.

29. B. Fischer, M. Hoffmann, H. Helm, G. Modjesch, P.U. Jepsen, Semicond. Sci. Technol. 20 (2005) S246.

30. Y. Li Hor, H.C. Lim, J.F Federici, E. Moore, J.W. Bozzelli, Chemical Physics 353 (2008) 185.

31. M. Walther, B.M. Fischer, P.U. Jepsen, PU. Chem. Phys. 288 (2003) 261.

32. Y. Watanabe, K. Kawase, T. Ikari, H. Ito, Y. Ishikawa, H. Minamide, Optics Communications 234 (2004) 125.

33. H. Wu, E.J. Heilweil, A.S. Hussain, M.A. Khan, Journal of Pharmaceutical Sciences 97 (2008) 970.

34. C.J. Strachan, P.F. Taday, D.A. Newham, K.C. Gordon, J.A. Zeitler, M. Pepper, T. Rades, Journal of Pharmaceutical Sciences 94 (2005) 837.

35. L. Ho, R. Müller, K.C. Gordon, P. Kleinebudde, M. Pepper, T. Rades, Y. Shen, P.F. Taday, J.A. Zeitler, Eur. J. of Pharm. and Biopharm. 71 (2009) 117.

36. C.M. McGoverin, T. Rades, K.C. Gordon, Journal of Pharmaceutical Sciences 97 (2008) 4598.

37. V. Malaterre, M. Perdesen, J. Ogorka, R. Gurny, N. Loggia, P.F. Taday, PF. Eur. J. Pharm. Biopharm. 74 (2010) 21.

38. L. Maurer, H. Leuenberger, International Journal of Pharmaceutics 370 (2009) 8.

39. A.J. Fitzgerald, B.E. Cole, P.F. Taday, J. Pharm. Sci. 94 (2005) 177.

40. L. Ho, R. Müller, K.C. Gordon, P. Kleinebudde, M. Pepper, T. Rades, Y. Shen, P.F. Taday, J.A. Zeitler, Journal of Controlled Release 127 (2008) 79.

41. R.P. Cogdill, R.N. Forcht, Y. Shen, P.F. Taday, J.R. Creekmore, C.A. Anderson, J.K. Drennen III, J. Pharm. Innov. 2 (2007) 29.

42. S.W. Smye, J.M. Chamberlain, A.J. Fitzgerald, E. Berry, E. Phys. Med. Biol. 46 (2001) R101.

43. B.M. Fischer, M. Walther, P.U. Jepsen, Phys. Med. Biol. 47 (2002) 3807.

44. Y.C. Shen, P.C. Upadhya, E.H. Linfield, A.G. Davies, Vibrational Spectroscopy 35 (2004) 111.

45. J.I. Nishizawa, T. Sasaki, T. Tanno, Journal of Physics and Chemistry of Solids 69 (2008) 693.

46. A.G. Markelz, A. Roitberg, E.J. Heilweil, Chem. Phys. Lett. 320 (2000) 42.

47. M.R. Kutteruf, C.M. Brown, L.K. Iwaki, M.B. Campbell, T.M. Korter, E.J. Heilweil, Chem. Phys. Lett. 375 (2003) 337.

48. M. Nagel, P.H. Bolivar, M. Bruscherseifer, H. Kurz, A. Bosserhoff, R. Büttner, Appl. Phys. Lett. 80 (2002) 154.

49. M. He, A.K. Azad, S. Ye, W. Zhang, Optics Communications 259 (2006) 389.

50. B. Ferguson, S. Wang, D. Gray, D. Abbott, X.C. Zhang, Microelectronics Journal 33 (2002) 1043.

51. V.P. Wallace, A.J. Fitzgerald, S. Shankar, N. Flanagan, R. Pye, J. Cluff, D.D. Arnone, British Journal of Dermatology 151 (2004) 424.

52. R.M. Woodward, V.P. Wallace, D.D. Arnone, E.H. Linfield, M. Pepper, Journal of Biological Physics 29 (2003) 257.

53. S.J. Oh, J. Kang, I. Maeng, J.S. Suh, Y.M. Huh, S. Haam, J.H. Son, Optics Express 15 (2009) 3469.

54. D.A. Crawley, C. Longbottom, B.E. Cole, C.M. Ciesla, D. Arnone, V.P. Wallace, M. Pepper, Caries Res. 37 (2003) 352.

55. M. Knott, New Scientist 2192 (1999) 22.

56. P.U. Jepsen, S.J. Clark, Chem. Phys. Lett. 442 (2007) 275.

2ème publication : Étude métrologique d'un spectromètre Terahertz : Faisabilité de la mesure de mélanges de poudres

Introduction

La première étape avant l'établissement de modèles de prédiction destinés à quantifier les produits est de contrôler efficacement l'acquisition des spectres. La plage spectrale utilisée correspond à la zone comprise entre 150 et 20 cm^{-1} car il s'agit des limites opérationnelles de la séparatrice que nous avons utilisée tout au long de la thèse mais surtout car il s'agit de la plage spectrale utilisée par le laboratoire Riken dont nous utilisons les spectres à fin de comparaison. Étant donné que nous nous sommes limités dans un premier temps à des expériences de caractérisation, nous n'avons pas désiré augmenter la plage de mesure pour nous concentrer sur les bandes et les pics caractéristiques connus des produits utilisés au cours de la thèse.

Cet article se focalise sur la caractérisation métrologique du spectromètre Terahertz sur lequel nous avons travaillé. Ne connaissant pas les caractéristiques ni le comportement précis du spectromètre, il a été nécessaire de procéder dans un premier temps à une étude métrologique sur des aspects simples : en l'occurrence la répétabilité de la mesure, la sensibilité de l'appareil et le rapport signal sur bruit. Il est nécessaire de connaître et de maîtriser ces paramètres avant de passer à une quelconque étude. Si des défauts ou des phénomènes particuliers apparaissent lors de l'acquisition des spectres, il faut pouvoir les reconnaître pour les retirer grâce à un prétraitement des spectres ou, faute de mieux, les prendre en compte lors des expériences pour en limiter l'impact.

Matériel et méthodes

Dispositif expérimental

Le spectromètre utilisé était un appareil de type « Bruker IFS 66v/S FTFIR », équipé d'une

séparatrice 23 µm en Mylar dont la gamme d'efficacité était comprise entre 150 et 20 cm^{-1} (0.6 - 4.5 THz). La source était une lampe au mercure émettant un rayonnement continu entre 600 et 5 cm^{-1} (0.15 – 18 THz). Le détecteur était un bolomètre au silicium de type « Infrared Laboratories Model N° HD-3 » refroidi à l'Hélium liquide à 4.2 K avec un filtre passe-haut, couvrant la gamme 0.5 - 10 THz. Le diaphragme présentait une ouverture de 12.0 mm. La résolution utilisée était de 0.5 cm^{-1}. La vitesse de balayage était fixé à 4 Hz. Toutes les mesures ont eu lieu sous vide pour éviter la présence des bandes caractéristiques de la vapeur d'eau et à température ambiante, 296 +/- 1 K.

La référence a été prise une seule fois, sans échantillon, au début des mesures. Chaque échantillon a ensuite été placé dans le spectromètre, sous vide, de façon à ce que le point de focalisation du rayon se situait au centre de la surface du comprimé faisant face à la source et son spectre a été acquis 125 fois, de manière automatique, successivement.

Les échantillons étaient constitués d'un mélange ternaire de poudres de polyéthylène (PE), saccharose et glucose. Le saccharose a été obtenu auprès de Fluka Analytical (Ref. N° 84100) et le glucose auprès de Sigma-Aldrich (Ref. N° 16325). Aucun n'a nécessité de purification supplémentaire. Le glucose présentait une granulométrie d'environ 100 µm et a été utilisé tel quel tandis que le saccharose a été broyé finement dans un mortier avant utilisation (granulométrie < 0.5 mm). Le PE a été acquis chez Aldrich (Ref N° 26935-2, spectrophotometric grade powder). La quantité de PE a été fixée à 70 % (m/m) de la masse totale des échantillons, hormis pour une pastille constituée de polyéthylène à 100 %. Les 30 % (m/m) restants ont été composés d'un mélange entre saccharose et glucose selon 21 pourcentages massiques : le premier échantillon contenait 0 % de glucose et 100 % de saccharose, le second 5 % de glucose et 95 % de saccharose, et ainsi de suite par pas de 5 % relatifs, jusqu'à 100 % de glucose. Les mélanges ont été ensuite compactés sous forme de pastilles de 0.5 ± 0.2 mm d'épaisseur et 12 mm de diamètre, pour un volume moyen de 56.55 mm^3. Chaque pastille a ensuite été re-broyée puis re-compactée, afin d'améliorer son homogénéité. La masse de chaque pastille était de 70 ± 0.1 mg.

Data Processing

Les spectres ont été collectés en transmission via le logiciel *Opus/IR* fourni avec le spectromètre puis traités sous Matlab version 7.4.0 (R2007a).

Un processus de détection des spectres aberrants a été réalisé sur chaque série de 125 spectres de la manière suivante : (i) Les 25 premiers spectres ont été retirés pour éliminer une éventuelle phase de transition au début de l'acquisition ; (ii) pour chacun des spectres restant, la distance de Mahalanobis à la moyenne de la série a été calculée, puis un test du T^2 de Hotelling a été réalisé avec un seuil de confiance de 1 %. Les individus dépassant ce seuil ont été considérés comme des individus aberrants ou outliers.

La totalité des spectres a été placée dans une matrice $\boldsymbol{T}$ de N lignes (individus : spectres) par P colonnes (variables : nombres d'onde), selon le schéma de la Figure 9 : les séries ont été regroupées en B blocs $\{\boldsymbol{T}^{1,}\cdots,\boldsymbol{T}^{k},\cdots,\boldsymbol{T}^{B}\}$ comprenant respectivement $\{N^{1,}\cdots,N^{k},\cdots,N^{B}\}$ spectres.

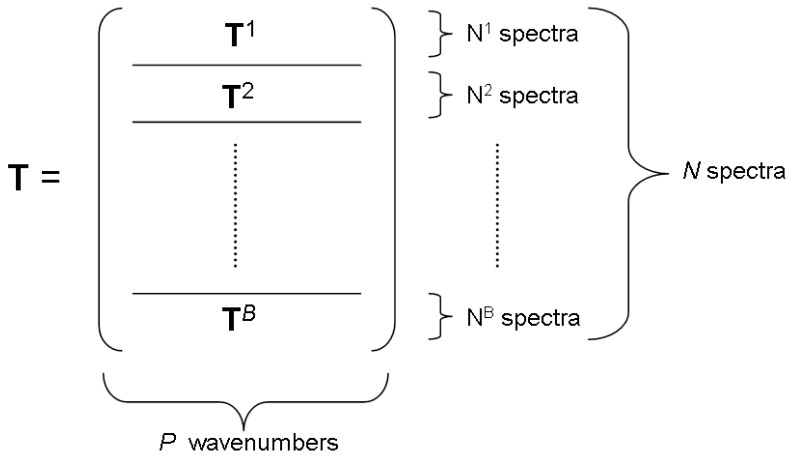

Figure 9: Matrice de regroupement des spectres en transmission.

La matrice $\boldsymbol{T}$ a été d'autre part convertie en absorbance (matrice $\boldsymbol{A}$) par la formule $\boldsymbol{A}=\log\frac{1}{\boldsymbol{T}}$ (Eq. 1). Les spectres en absorbance ont ensuite été moyennés par bloc pour obtenir la matrice $\boldsymbol{A}_{\mathrm{m}}$.

Les pourcentages en glucose et saccharose ont été regroupés dans le Tableau 1 ; un bloc de mesure nommé $\mathbf{T}^0$ a également été acquis et a été constitué de 100 scans d'un comprimé de PE pur.

Bloc de mesure	% Glucose	% Saccharose
$\mathbf{T}^{1}$	0	100
$\mathbf{T}^{2}$	5	95
$\mathbf{T}^{3}$	10	90
$\mathbf{T}^{4}$	15	85
$\mathbf{T}^{5}$	20	80
$\mathbf{T}^{6}$	25	75
$\mathbf{T}^{7}$	30	70
$\mathbf{T}^{8}$	35	65
$\mathbf{T}^{9}$	40	60
$\mathbf{T}^{10}$	45	55
$\mathbf{T}^{11}$	50	50
$\mathbf{T}^{12}$	60	40
$\mathbf{T}^{13}$	65	35
$\mathbf{T}^{14}$	70	30
$\mathbf{T}^{15}$	75	25
$\mathbf{T}^{16}$	80	20
$\mathbf{T}^{17}$	85	15
$\mathbf{T}^{18}$	90	10
$\mathbf{T}^{19}$	95	5
$\mathbf{T}^{20}$	100	0

Tableau 1 : Composition en sucres (en pourcentage relatif, polyéthylène non inclus) des blocs de mesure.

Notez que les séries $\mathbf{T}^{11}$ et $\mathbf{T}^{12}$ sont passées de 50 à 60 % de glucose, sans comporter d'échantillon à 55 % de glucose et 45 % de saccharose. Ceci a été dû au fait que l'échantillon originel s'est brisé avant la mesure et a été contaminé, son remplacement ayant été impossible à cause de la pénurie en polyéthylène de qualité spectroscopique (ce dernier est le seul valable pour les mesures dans nos conditions expérimentales et il a été impossible d'en trouver plus de stock, n'étant plus distribué depuis peu).

Analyse univariée

Le rapport signal sur bruit (SNR) univarié a été évalué pour chaque nombre d'onde w_j sur la série du

PE, selon la formule de l'équation 2 : $SNR_{mono}(w_j)=\frac{E(\boldsymbol{T}_j^0)}{\sigma(\boldsymbol{T}_j^0)}$, où $\boldsymbol{T}^0$ était la matrice contenant les spectres, avec $\sigma(\boldsymbol{T}_j^0)$ et $E(\boldsymbol{T}_j^0)$ respectivement l'écart-type et l'espérance mathématique de la colonne *j* de $\boldsymbol{T}^0$.

La répétabilité en nombre d'onde a été évaluée en calculant la position de l'apex d'une des bandes caractéristiques du glucose (à 48 cm^{-1}) pour chaque spectre de la série $\boldsymbol{T}^{20}$ correspondant au mélange 70 % PE – 30 % G (m/m), puis en calculant l'écart type σ_w de cette valeur.

La répétabilité en transmission univariée a été évaluée par l'écart-type intra blocs de chaque colonne de $\boldsymbol{T}^k$, selon la formule suivante :

$$R_{mono}(w_j)=\sqrt{\frac{\sum_{k=1}^{B}\sum_{i=1}^{N^k}(t_{ij}^k-E(\boldsymbol{T}_j^k))}{N-B}} \quad \text{(Eq. 3)}$$

La sensibilité univariée $S_{mono}(w_j)$ a été évaluée par la pente de la relation entre chaque colonne de $\mathbf{A_m}$ (matrice centrée) et la concentration volumique en glucose **y**.

Analyse multivariée

Le rapport signal sur bruit multivarié a été évalué par le rapport de la norme du spectre du PE sur sa variance. Ce calcul a été effectué de la manière suivante. Le spectre moyen du PE $\boldsymbol{t}^0$ a été calculé par $\boldsymbol{t}^0=(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\boldsymbol{T}^0$ (Eq. 4) puis la matrice $\boldsymbol{T}^0$ a été centrée par $\boldsymbol{T}_C^0=\boldsymbol{T}^0-\mathbf{1}\boldsymbol{t}^0$ (Eq. 5). Le rapport signal sur bruit multivarié a été évalué par le ratio suivant :

$$SNR_{multi}=\frac{\sqrt{\boldsymbol{t}^{0\mathrm{T}}\boldsymbol{t}^0}}{\sqrt{trace(\frac{1}{N^0}\boldsymbol{T}_C^{0\mathrm{T}}\boldsymbol{T}_C^0)}} \quad \text{(Eq. 6).}$$

À partir de la matrice $\boldsymbol{T}$ et des blocs $\boldsymbol{T}^k$, les matrices de variance co-variance totale (***S***), intra blocs (***W***) et inter blocs (***B***) ont été calculées de la manière suivante :

$\boldsymbol{S} = \frac{1}{B-1}\boldsymbol{Var}(\boldsymbol{T})$ (Eq. 7) **;**

$\boldsymbol{W} = \frac{1}{N-B}\sum_{k}(N^k - 1)\boldsymbol{Var}(\boldsymbol{T^k})$ (Eq. 8) ;

$\boldsymbol{B} = \boldsymbol{S} - \boldsymbol{W}$ (Eq. 9) ;

avec $\boldsymbol{Var}(\boldsymbol{T^k})$ la matrice de variance co-variance du bloc de mesure *k*, $\boldsymbol{g}$ le centre de gravité du nuage de points global et $\boldsymbol{g}^k$ les centres de gravité des blocs $\boldsymbol{T^k}$.

La répétabilité en transmission multivariée a été caractérisée par l'analyse de la matrice ***W***. La sensibilité multivariée a été caractérisée par l'analyse de la matrice ***B***. Les éléments propres (valeurs propres et vecteurs propres) de ces matrices expliquant au moins 80 % de variance ont été calculés et analysés.

Les spectres nets (Net Analyte Signal) des deux composés glucose et saccharose ont été extraits des spectres d'absorbance par la formule de l'équation 10 : $\boldsymbol{K} = (\boldsymbol{Y}^T\boldsymbol{Y})^{-1}\boldsymbol{Y}^T\boldsymbol{A}_m$, avec ***Y*** la matrice des concentrations des sucres exprimées en $g.L^{-1}$, et $\boldsymbol{A}_m$ la matrice des spectres moyens de chaque série. Conformément aux études de Faber [Faber 1999] la sensibilité multivariée du glucose et du saccharose ont été évaluées par la norme des NAS, i.e. des deux lignes de ***K***.

Présentation de l'article

Cet article est également l'occasion de comparer les méthodes d'analyses spectrales basées sur des techniques univariées avec des techniques multivariées. Le lectorat auquel est destiné cette publication est un public de spectroscopistes, parfois peu spécialisés dans les techniques multivariées.

Discussion

Les spectres des produits que nous avons choisi d'étudier, le glucose et le saccharose, sont disponibles dans la base de données en ligne de Riken. Nous avons sélectionné ces sucres en vue de comparer, dans un premier temps, nos spectres expérimentaux avec ceux de Riken. Il s'est avéré

qu'il y a des différences entre les spectres, comme le montre la Fig 2 de la publication à suivre. Nous n'avons malheureusement pu formuler que des hypothèses au sujet de ces différences puisqu'il n'y a pas d'information disponible sur le site concernant leur méthodologie d'acquisition ou le matériel utilisé et nous n'avons pas pu obtenir d'informations auprès des membres de Riken avant la fin de nos travaux. Nous pouvons donc supposer que les différences entre les spectres proviennent à la fois de la source (laser à cascade quantique, lampe mercure ?), du détecteur employé, de la préparation des comprimés (masse, pression appliquée, présentation de l'échantillon) ainsi que des conditions expérimentales. Tout ceci montre bien que les différences d'appareillage conduisent à des problèmes de reproductibilité entre notre étude et la leur. Les spectres Riken ont donc été uniquement utilisés en termes de comparaison d'emplacement des bandes caractéristiques des sucres.

L'intérêt de cet article porte principalement sur la comparaison des approches univariées et multivariées en ce qui concerne le traitement des spectres. Des distinctions importantes peuvent être établies entre ces deux approches :

- Univariée :
    - Unique variable de concentration, signal mesuré unique.
    - Régression linéaire standard linéaire ou non-linéaire univariée.
    - Données expérimentales entièrement représentées par un unique point, bi-dimensionnel.
- Multivariée :
    - k variables de concentration, k signaux mesurés.
    - Régression multivariée (plusieurs approches).
    - Visualisation complexe des données expérimentales.

Dans les analyses univariées, la variable mesurée doit être fortement sélective pour être d'intérêt.

Par exemple, si seulement un composé d'un mélange absorbe à une longueur d'onde particulière, alors l'absorbance du mélange à cette longueur d'onde est proportionnelle (en assumant que la loi de Beer-Lambert est respectée) à la concentration du composé. S'il y a deux composés absorbants, alors l'absorbance mesurée à cette longueur d'onde n'est plus proportionnelle à la première concentration seulement.

S'il y a plusieurs espèces absorbantes, il y a deux façons de procéder. La première est de restaurer la sélectivité de la méthode univariée en séparant physiquement le composé à mesurer des autres espèces (par chromatographie par exemple). La seconde est de réaliser une série de mesures sélectives (comme la mesure de l'absorbance à un certain nombre de longueurs d'onde) et utiliser une régression multivariée pour trouver une combinaison linéaire des mesures qui est totalement sélective vis-à-vis du composé.

La première de ces options est celle qui consume le plus de temps et éventuellement de ressources. En revanche la seconde est plus complexe d'un point de vue mathématique et constitue le premier aspect de la définition de la chimiométrie selon Wold donnée dans la partie Introduction de ce manuscrit, mais est au final beaucoup plus rapide. Le travail effectué pour cette publication tient à comparer ces deux approches et montrer les avantages de l'approche multivariée par rapport à l'approche univariée.

L'analyse des spectres a montré que la quasi-totalité des individus aberrants sont des spectres très fortement bruités. Ceci est dû à l'extrême sensibilité du détecteur. En effet, un claquement de porte, une mise sous tension d'un appareil ou d'une bobine électromagnétique peut provoquer des interférences qui se répercuteront sur le spectre. Il est également intéressant de noter que les spectres aberrants ne se suivent pas au sein d'une même série, ils sont aléatoires, ce qui montre qu'il s'agit donc de phénomènes isolés lors de la mesure, peu probablement liés à l'appareillage.

Le fait que le premier loading soit le spectre du saccharose plutôt que la relation entre la concentration en sucres et l'absorbance (façon dont a été déterminée la sensibilité en analyse univariée) indique que le spectromètre discerne beaucoup mieux les variations en saccharose. Le glucose ne fournit pas une réponse très importante à première vue. Comme nous l'avons vu lors de l'étude de la sensibilité univariée, la loi de Mie indique que la granulométrie a un effet direct sur l'absorbance d'un composé. D'après notre hypothèse, l'analyse multivariée montre que la taille des grains a un effet direct sur la détection des produits et sur leur caractérisation par le spectromètre. Le saccharose est beaucoup mieux reconnu par l'appareil, au détriment du glucose.

Les spectres sont projetés dans un plan à deux dimensions constitués par les loadings, plan qui génère la matrice ***B***. Dans notre cas, la loi de Beer-Lambert peut s'écrire de différentes manières. A longueur de trajet équivalente, l'absorbance d'un spectre *x* peut se calculer en faisant la somme des concentrations en sucres multipliés respectivement par leurs coefficients d'absorption molaire comme le montre l'équation 1.

$$x = C_S \varepsilon_S + C_G \varepsilon_G \quad \text{Eq. 1}$$

En développant l'équation, étant donné qu'en exprimant les concentrations en $g.L^{-1}$ on peut exprimer les concentrations l'une en fonction de l'autre, il est possible d'exprimer l'absorbance d'un spectre uniquement en fonction d'une concentration, celle du glucose, et des coefficients d'absorption, la formule devenant alors :

$$x = \varepsilon_S + (\varepsilon_G - \varepsilon_S) C_G \quad \text{Eq. 2}$$

Ces coefficients peuvent être représentés par les loadings. 80 % de la variance est expliquée par le premier loading, soit le spectre du saccharose. A ce stade, nous avons formulé l'hypothèse que le saccharose est mieux perçu par le spectromètre que le glucose. Le deuxième loading étant orthogonal au premier par construction, il correspond à la différence entre le spectre de glucose et le spectre de saccharose. Cela correspond au spectre calculé pour la sensibilité univariée exprimée en fonction du saccharose illustré Figure 10. Nous retrouvons bien les deux loadings de ***B*** dans cette

formule.

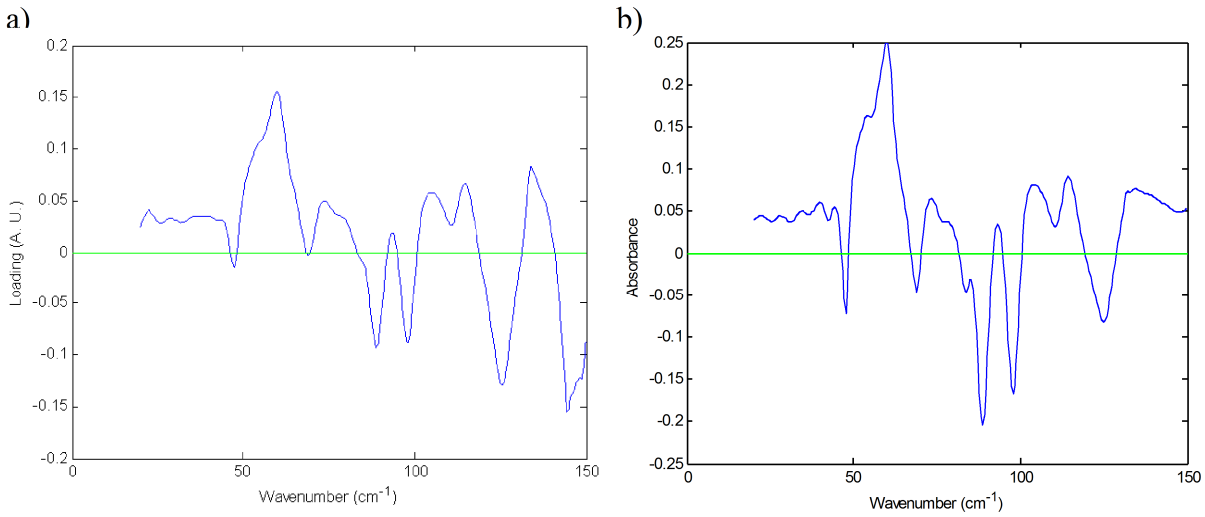

*Figure 10: a) Deuxième loading de **B** ; b) Différence entre les spectres de glucose et de saccharose.*

Il faut toutefois noter que nous ne savons pas si cette explication est valable pour tous les mélanges de poudres, quels que soient les produits utilisés ou s'il ne s'agit que d'une heureuse coïncidence. S'il ne fait aucun doute que la loi de Beer-Lambert peut être appliquée à la spectroscopie THz, il sera nécessaire de procéder à d'autres expériences avec d'autres produits afin de confirmer ou d'infirmer cette loi de proportionnalité qui fonctionne très bien dans notre cas.

Conclusion

Cet article montre les avantages à utiliser des techniques multivariées plutôt que des analyses univariées lors d'une étude spectrale. Elles permettent en effet de faire ressortir exactement les phénomènes qui ont lieu lors de l'acquisition, et permettent leur explication de manière plus précise qu'avec un point de vue univarié. Cela est principalement expliqué parce que la technique d'Analyse en Composantes Principales que nous avons utilisée ici tient compte que des variables spectrales sont corrélées les unes avec les autres tandis que des techniques univariées ne se concentrent que sur une seule variable à la fois. Ainsi, des structures, des bruits internes aux spectres peuvent être mis en évidence et on peut agir dessus pour tenter de les retirer ou de

minimiser leurs effets. Ici, nous avons fait l'hypothèse que le bruit sinusoïdal est caractéristique d'un décalage de l'interférogramme. Il devrait être possible de le corriger par des méthodes chimiométriques en retirant la fréquence responsable du sinus, de projeter orthogonalement la matrice de données à celle du sinus afin d'en retire l'effet ou en décalant l'interférogramme entier, mais cela peut se traduire par des modifications sur les données spectrales reliées aux informations caractéristiques des produits. Néanmoins, les tentatives effectuées pour diminuer l'erreur de répétabilité via une correction de l'interférogramme n'ont pas donné satisfaction : soit il restait une sinusoïde (à cause de la complexité du bruit et du fait qu'il soit porté par les deux premiers loadings au moins), soit des informations liées aux produits étaient retirées lors de la correction. Il est donc conseillé de travailler avec la plus grande circonspection quand ces solutions sont choisies. L'autre solution consiste à intervenir directement sur le spectromètre. C'est ce qui a eu lieu dans notre situation. L'examen de l'appareil a révélé que la butée du miroir situé sur le Michelson n'était plus fixée à sa place. En conséquence, le miroir parcourait une distance légèrement différente lors des mesures. Après intervention, nous avons pu constater que le sinus avait disparu lors de l'acquisition des spectres. Il a donc été décidé, pour la suite des travaux, de conserver directement le spectre moyen des scans d'une mesure.

Ces expériences ont été une préparation nécessaire en vue de la prochaine étape de notre recherche : la quantification et la distinction de produits pulvérulents dans des mélanges.

Publication

*as submitted in **Applied Spectroscopy**.*

METHODOLOGICAL STUDY OF A TERAHERTZ SPECTROMETER: FEASIBILITY OF POWDER MIXTURE ASSESSMENTS

M. Papillaud*[a], C. Gergely[b,c], F. Davrieux[a], J.M. Roger[d]

[a] Centre International pour la Recherche Agronomique et le Développement (CIRAD), UMR 95 Qualisud, TA B-95/16, 73 rue J. F. Breton 34398 Montpellier Cedex 5, France

[b] Université Montpellier 2, Laboratoire Charles Coulomb UMR 5221, F-34095, Montpellier, France

[c] CNRS, Laboratoire Charles Coulomb UMR 5221, F-34095 Montpellier, France

[d] UMR ITAP, Cemagref, BP 5095, Montpellier Cedex 1, France

*Corresponding author: Matthieu Papillaud, CIRAD UMR Qualisud, TA B-95/16, 73 rue J. F. Breton 34398 Montpellier Cedex 5, France. Phone: +33(0)4 67 61 54 52; Fax: +33(0)4 67 61 44 33. Email: mpapillaud@yahoo.fr

Abstract

Development of new analytical techniques requires to control or at least to know various device parameters, such as repeatability and sensitivity. Terahertz spectroscopy has proven to be a new asset in the analytical range, but little is known about the insight behavior of spectra. Method repeatability and spectrometer sensitivity have been studied through a comparison between univariate and multivariate approaches. In this study, we have shown that multivariate analysis allows complementing univariate analysis and gives more information concerning spectral behavior. Unseen features and spectral noise, such as sinusoid shaped noise applied to spectra, undetected and unexplained with univariate analysis, are revealed and decomposed thanks to multivariate techniques. They also have shown that contrary to what preliminary univariate analysis suggested, both products are equally detected by our spectrometer. This proves the necessity to work with multivariate methods to be aware of phenomena unrelated to spectral data, in order to control or eventually remove them before any further study.

Keywords

Terahertz spectroscopy (THz), multivariate analysis, net analyte signal (NAS), univariate analysis, metrology, repeatability, sensitivity.

Introduction

Terahertz (1 THz = 10^{12} Hz) technology is a continuously developing technique since the 1980s.[1] Constant innovation in the semi-conductors area guaranteed the success of this technique through the development of brighter sources and detectors that are more sensitive to THz waves and user-friendly. Many domains, including chemistry, biology and pharmaceutical industry, are using THz radiation to fill the gaps in various spectroscopic domains or complement other known and well-used techniques such as infrared spectroscopy.

THz frequency domain covers approximately the region from 100 GHz to 30 THz, i.e. the wavenumbers between 3 and 1000 cm^{-1}. Many materials have a unique spectral fingerprint in this region whereas some others can be transparent or semi-transparent (textiles, paper, wood, cardboard, plastics…). THz rays are weakly energetic and non-ionizing which renders them harmless. The strong absorption of THz waves by liquid water suggests a good interaction between biological samples and THz waves. These waves induce rotation and vibration of polar water molecules and excite low energy intermolecular bonds, such as hydrogen bonds, in water and proteins for example. Biological domain uses actively THz spectroscopy to study the DNA hybridation state[2] or to collect the spectra of numerous biomolecules as it has been reviewed by Plusquellic.[3] Sensitive to intermolecular vibrations, THz spectroscopy provides information related to structure and immediate environment of analyzed molecules. Taday published results of industrial interest when clearly differentiating three isomers of the paracetamol,[4] and characterizing the amorphous or the crystalline form of a molecule.[5] These molecules are characterized

by very different THz spectra, allowing quick recognition with no complementary analysis required which is a great challenge for industrials. The interest for THz spectroscopy led to the identification and the characterization of several chemical products such as drugs,[6, 7] explosives,[6, 8-11] pesticides[12] and sugars[13, 14] among others.

Most of these works focus on product identification and data collection, sometimes on the prediction of characteristic bands location with the help of previous results obtained with other spectroscopic techniques.[7-12] A few studies, mostly related to the pharmaceutical domain,[4, 5] use their data to develop prediction or quantification models. One way to construct such models is to use univariate analysis, analyzing spectra wavelength after wavelength, and eventually selecting some wavelengths of interest. According to the uses in near infrared spectroscopy for instance, multivariate analysis techniques, taking in consideration the whole spectral zone, permit to analyze the spectra allure and find the information in structural patterns (peaks, inflexion points, bands…).[15, 16] Moreover, in order to create prediction models reliable in time, parameters such as the repeatability and the sensitivity of the spectrometer must be controlled or at least known to ensure the quality of the experiments and obtained results. This paper proposes to realize the metrological study of a THz spectrometer on ternary mixtures of sugars and polyethylene (PE) and to demonstrate the benefits of a multivariate analysis with regard to a univariate analysis.

Notation

Capital bold characters will be used for matrices, e. g. **X**; small bold characters for column

vectors, e. g. $\mathbf{x}_i$ will denote the i^{th} column of **X**; row vectors will be denoted by the transpose notation, e. g. $\mathbf{x}^T_j$ will denote the j^{th} row of **X**; column vector transpose will be denoted $(\mathbf{x}_i)^T$, italic characters will be used for scalars, e. g. matrix elements x_{ij} or indices *i*. From a given matrix **X** of *N* rows and *P* columns, the mean row, denoted by $\bar{x}$, is calculated by $\bar{x}^T = (\mathbf{1}_N^T \mathbf{1}_N)^{-1} \mathbf{1}_N^T X_C$ where $\mathbf{1}_N$ is a column vector of *N* ones and the centered matrix will be denoted X_C and given by $X_C = X - \mathbf{1}_N \bar{x}^T$.

Material and Methods

Experimental setup

The spectrometer was a “Bruker IFS 66v/S FTFIR”, equipped with a 23µm Mylar beamsplitter which efficacy range covered from 150 to 20 cm^{-1} (0.6 – 4.5 THz). The source was a Hg lamp emitting a continuous radiation between 600 and 5 cm^{-1} (0.15 – 18 THz). The detector was a Si bolometer from “Infrared Laboratories Model” N° HD-3 which was cooled with liquid He at 4.2 K. It was equipped with a high-pass filter which covered the range between 0.5 to 10 THz. The diaphragm had an aperture of 12.0 mm wide. The resolution used has been set to 0.5 cm^{-1}. All measurements were under vacuum and at ambient temperature, 296 +/- 1 K. Reference has been registered once, without any sample, at the beginning of the experiments. Each sample has been measured by collecting 125 individual scans automatically.

Samples were made up of ternary mixtures of PE powder, sucrose and glucose. Sucrose was obtained from Fluka Analytical (Ref. N° 84100) and glucose was purchased from Sigma-Aldrich (Ref. N° 16325). No further purification was needed for both products.

Glucose powder of size <100 µm was used with no other preparation whereas sucrose has been grounded with a mortar and a pestle to reduce particle size (<500 µm). PE powder of size 100 µm was purchased from Aldrich (Ref. N° 26935-2, spectrophotometric grade powder). Sample discs were prepared by mixing different ratios of sugars with PE. The quantity of PE in discs was fixed at 70 % (m/m) of the total mass. The remaining 30 % (m/m) were composed of a mixture of glucose and sucrose, for a total of 21 samples. The first sample contained 0 % of glucose and 100 % of sucrose, the second 5 % of glucose and 95 % of sucrose and so on, with 5 % of mass changes, up to 100 % sucrose. The mixtures have been compacted in 0.5 ± 0.2 mm thick and 12 mm wide discs with a mean volume of 56.55 mm^3. Each disc has been grounded and compacted a second time to ensure homogeneity. Disc masses were of 70 ± 0.1 mg.

Data processing

Spectra were collected in transmission with the *Opus/IR* software and pretreated with Matlab (version 7.4.0 R2007a).

Detection of outliers was conducted for each set of 125 spectra with this method: (i) the first 25 spectra were removed in order to avoid any transition phase that should occur at the beginning of the acquisition; (ii) for each one of the remaining spectra the Mahalanobis distance to the mean sample has been calculated, then a Hotelling T^2 test has been performed with a confidence level of 1 %. Samples that exceeded this level were considered as outliers.

The remaining transmission spectra were arranged in a matrix **T** of N rows (individual: spectrum) by P columns (variable: wavenumber), as illustrated in Fig 1: sets have been

gathered in B blocks {$\mathbf{T}^1$, ..., $\mathbf{T}^k$, …, $\mathbf{T}^B$} each one containing {N^1, …, N^k, …, N^B} spectra respectively.

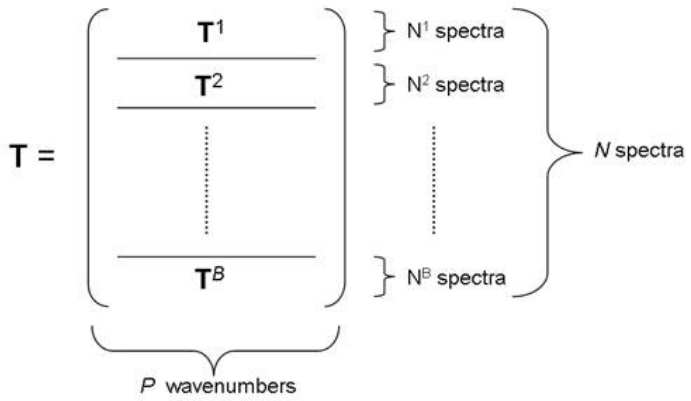

Fig 1: Global matrix T regrouping transmission spectra.

Matrix **T** has then been converted to the absorbance matrix **A** with the equation $\boldsymbol{A}=\log\frac{1}{\boldsymbol{T}}$ (Eq. 1). The mean spectrum of each block has then been calculated and put into matrix $\mathbf{A_m}$. One additional block $\mathbf{T}^0$ has been created with the scans pf a pure polyethylene pellet.

Univariate analysis

The univariate signal to noise ratio (SNR) has been estimated for each wavenumber w_j of a pure PE sample measurements using Eq. 2:

$$SNR_{mono}(w_j)=\frac{E(\boldsymbol{T}_j^0)}{\sigma(\boldsymbol{T}_j^0)} \quad \text{(Eq. 2)}$$

With $\mathbf{T}^0$ the matrix with the PE spectra and $\sigma(\boldsymbol{T}_j^0)$ and $E(\boldsymbol{T}_j^0)$ the mean and the standard deviation respectively of the *j* column of $\mathbf{T}^0$.

The wavenumber repeatability error has been estimated by determining the value of the apex of one of the characteristic features of glucose (at 48 cm^{-1}) for each spectrum of the $\mathbf{T}^{20}$ series, which corresponded to the 70 % PE – 30 % glucose mixture (m/m), then by

calculating the standard deviation σ_j for this position.

The transmission repeatability error has been estimated by the square root of the within group scatter for each column of $\mathbf{T}^k$, according to Eq. 3:

$$R_{mono}(w_j)=\sqrt{\frac{\sum_{k=1}^{B}\sum_{i=1}^{N^k}(t_{ij}^k-E(\boldsymbol{T}_j^k))^2}{N}} \quad \text{(Eq. 3)}$$

univariate sensitivity $S_{mono}(w_j)$ has been estimated by the slope between each column of $\mathbf{A_m}$ (centered matrix) and the volumetric concentration of sugars.

Multivariate analysis

The mean spectra from PE $\boldsymbol{t}^0$ has been calculated by $\boldsymbol{t}^0=(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\boldsymbol{T}^0$ (Eq. 4) then the $\boldsymbol{T}^0$ matrix has been centered with $\boldsymbol{T}_C^0=\boldsymbol{T}^0-\mathbf{1}\boldsymbol{t}^0$ (Eq. 5).

Multivariate SNR has been evaluated by the following equation (Eq. 6):

$$SNR_{multi}=\frac{\sqrt{\bar{\boldsymbol{t}}^{\mathbf{0}T}\bar{\boldsymbol{t}}^{\mathbf{0}}}}{\sqrt{trace\left(\frac{1}{N^0}\boldsymbol{T}_C^{0\mathrm{T}}\boldsymbol{T}_C^0\right)}} \quad \text{(Eq. 6)}$$

From matrix **T** and $\mathbf{T}^k$ blocks, the following scatter matrices: total (**S**), within (**W**) and between (**B**), have been calculated as follows:

$$\boldsymbol{S}=\frac{1}{N}\boldsymbol{T}_C^T\boldsymbol{T}_C \quad \text{(Eq. 7)}$$

$$\boldsymbol{W}=\frac{1}{N}\sum\boldsymbol{T}_C^{kT}\boldsymbol{T}_C^k \quad \text{(Eq. 8)}$$

$$\boldsymbol{B}=\boldsymbol{S}-\boldsymbol{W} \quad \text{(Eq. 9)}$$

Multivariate repeatability has been characterized in transmission by the **W** matrix analysis.

Multivariate sensitivity has been characterized by the **B** matrix analysis. Matrices eigenvectors and eigenvalues explaining at least 80 % of the total variance have been calculated and analyzed.

Net spectra (Net Analyte Signal, NAS) of the two products glucose and sucrose have been extracted from absorbance spectra thanks to Eq. 10:[15]

$$\mathbf{K} = (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{A_m} \quad \text{(Eq. 10)}$$

With **Y** the matrix regrouping the sugars concentration of $\mathbf{T}^k$ blocks. According to the work of Faber,[17] multivariate sensitivity of the spectrometer to glucose and sucrose has been evaluated by the norm of NAS, i.e. **K** rows.

Results and Discussion

Figure 2 compares our experimental THz spectra of pure glucose and pure sucrose and spectra acquired from the online THz spectral database of Riken [From THz database Web: http://www.riken.jp/THzdatabase/ (Tera-photonics Laboratory, RIKEN Sendai)]. Riken spectra correspond to samples acquired as discs mixing sugar and PE which has been chosen as optical matrix because of its high transparency to THz rays.

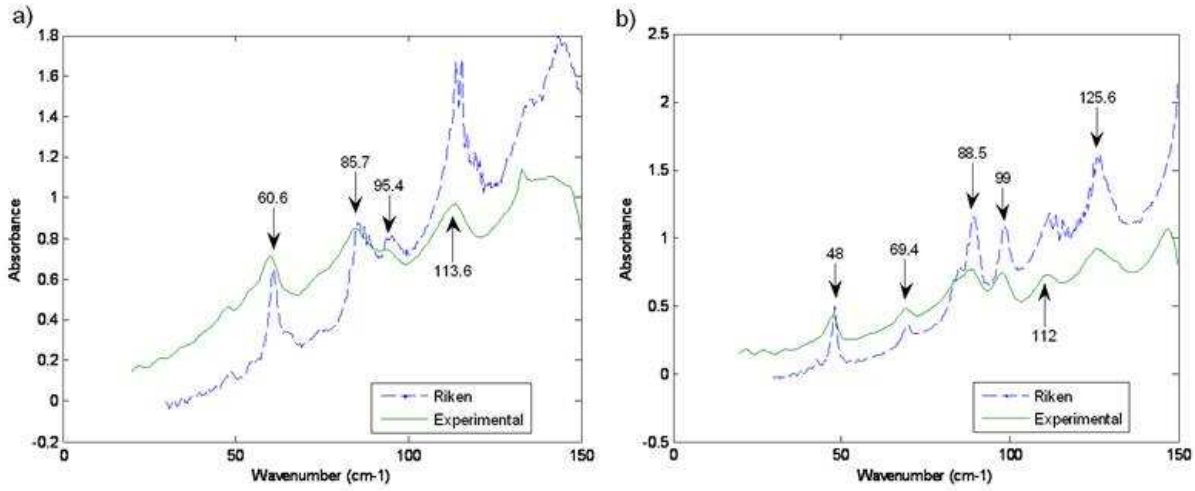

Fig 2: a) sucrose absorbance spectra, b) glucose absorbance spectra.

Both spectra, experimental as well as the Riken ones, present a strong baseline. This artifact is probably due to photon diffusion in discs. One can also see that baseline is more important for Riken spectra than for our experimental spectra. This might be the consequence of the difference in the granulometry in samples which was not completely controlled and is known as the principal source of diffusion.

It also appears that experimental spectra are smoother than Riken spectra. This difference exists at two levels:

- (i) Riken spectra present noise at high frequency, for example next to 120 cm^{-1} for glucose, contrary to experimental spectra. This might be explained by the difference of the used resolution: 0.2 cm^{-1} for Riken and 0.5 cm^{-1} for our experiments.

- (ii) Global features (principally peaks) are weaker in experimental spectra than in Riken spectra. Peaks net height (after baseline removal) is almost two to three times more important for Riken's spectra. Their samples have been measured with 2 mm thick discs at a concentration of 10 % (m/m). Our experimental samples have a four times less thickness and a three times higher concentration. This would explain a ¾ ratio between peaks heights. The rest of the difference is certainly linked to the used instruments (resolution)

and to the experimental setup (it is unknown whether database reference has been taken in vacuum or with a pure PE disc).

The characteristic peaks of glucose inventoried in Riken database (48, 69.4, 88.5, 99 and 125.6 cm^{-1}) can be identified in the experimental spectrum. An additional absorbance peak appears in the experimental spectrum at 112 cm^{-1} which is not referenced in the Riken database, most certainly because it is located in a noisy area. Another peak at 147 cm^{-1} is to be noted in experimental spectrum, while it does not appear on Riken spectrum.

The characteristic peaks of sucrose identified by Riken (60.6, 85.7, 95.4 and 113.6 cm^{-1}) have also been observed in our experimental spectrum. A weak absorbance peak at 48 cm^{-1} is more revealed in the experimental spectrum than in Riken's spectrum.

Figure 3 collects the mean spectra of the following absorbance series: $\mathbf{A}^{1}$, $\mathbf{A}^{5}$, $\mathbf{A}^{9}$, $\mathbf{A}^{12}$, $\mathbf{A}^{16}$, $\mathbf{A}^{20}$.

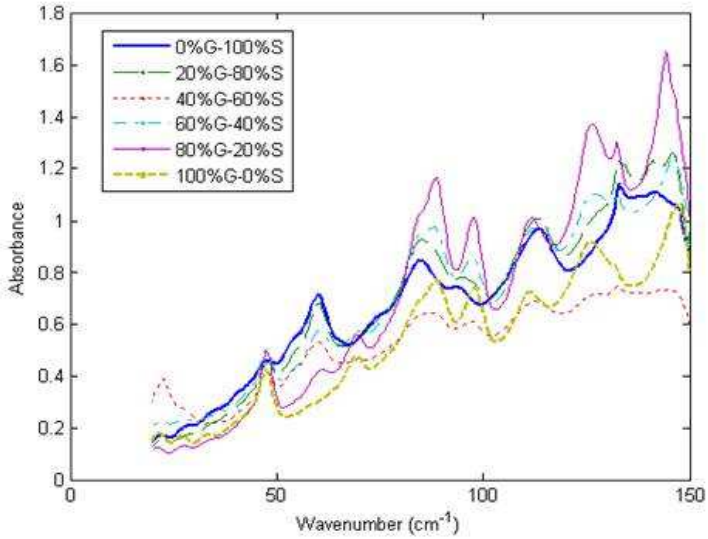

Fig 3: Mean spectra in absorbance from six samples blocks.

One can identify again the characteristic features of glucose and sucrose shown in Figure 2. This suggests that there has not been any interaction between the products, which consequence would have been to shift peaks.

We note the apparition of a common band to sucrose and glucose spectra at 47.4 cm^{-1}. The combination of the two peaks at 99 cm^{-1} for glucose and at 95.4 cm^{-1} for sucrose creates a large feature. Features that appear beyond 130 cm^{-1} do not belong to the products. These measurements artifacts are caused by the beamsplitter high absorption at its lower and higher limits. This implies that the zone under 30 cm^{-1} is not really reliable. Finally, one can see that the mixture baselines are not identical for each spectrum: they are neither constant nor in order. Low homogeneity in samples could stand as an explanation. The consequence of these different baselines is that spectra cross themselves without getting along with the glucose or sucrose concentration. Nevertheless at 60 cm^{-1}, where a sucrose peak can be found, peaks height vary in satisfying manner compared to the concentration in sucrose.

Univariate analysis

Wavenumber repeatability

For each of the $\mathbf{A}^{20}$ spectra, maximum absorbance has been found at the same exact wavenumber: 47.4 cm^{-1}. It appears that the wavenumber repeatability error of the apparatus is lower than its resolution.

Signal to Noise Ratio

Fig 4 a) shows the 23 µm beamsplitter efficiency zone. As the spectrum is recorded in transmission, it appears that the zone included between 50 and 120 cm^{-1} yields a better transmission. On the contrary, the extremities of the measurement range, before 50 cm^{-1} and beyond 130 cm^{-1}, show that the beamsplitter is less transparent. As a result, spectral information or features found in these areas can be dwarfed or not seen. Lower signal around the extremities also suggest that SNR might be weak and bother the experimental

spectra interpretation. It might be the explanation of the apparition of the peak at 147 cm^{-1} in our experimental glucose spectrum, which could result from the combination of the peak base at 153 cm^{-1} and the higher limit of the beamsplitter. Similarly, for the sucrose spectrum, our experimental spectrum shows beyond 130 cm^{-1} an artifact that is not characteristic of sucrose, but is caused by the higher limit of the beamsplitter.

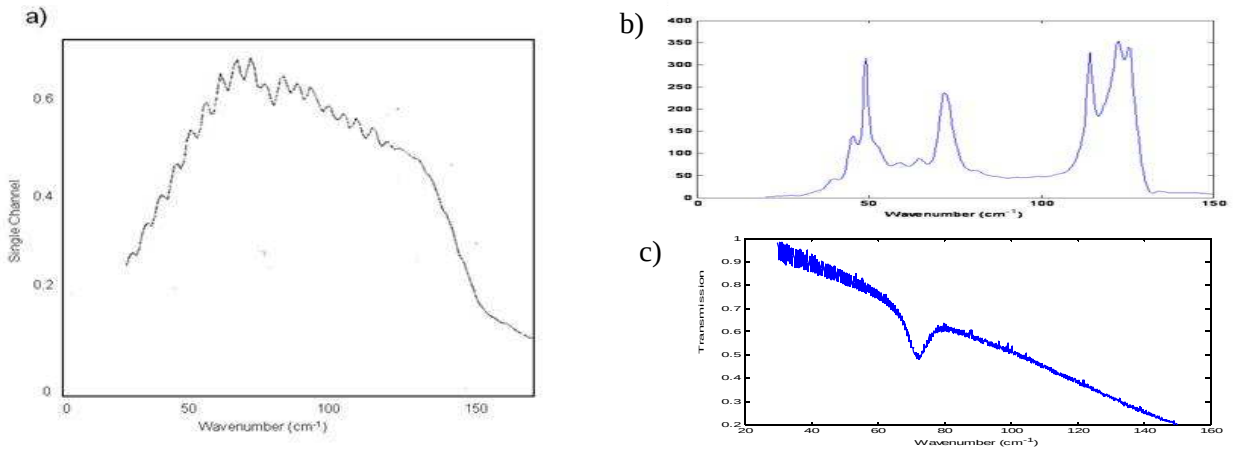

Fig 4: a) Beamsplitter efficiency spectrum in transmission (manufacturer data); b) Univariate SNR calculated with pure PE spectra; c) PE transmission spectrum.

SNR is shown Figure 4 b) as it has been calculated for the 20 – 150 cm^{-1} range. Figure 4 c) represents the polyethylene spectrum, with a single characteristic band at 71 cm^{-1}. SNR is neither uniform nor constant on the whole range. Areas around 50, 75 and from 110 to 130 cm^{-1} present strong coefficients. Transmission is not very influenced by noise coming from the apparatus itself. On the contrary, the spectral range extremities, 20 – 30 cm^{-1} and 130 – 150 cm^{-1}, present low values for SNR where noise is more important during the signal acquisition. Therefore, range between 40 and 130 cm^{-1} would be more effective for measurements, regarding the high SNR values.

Repeatability

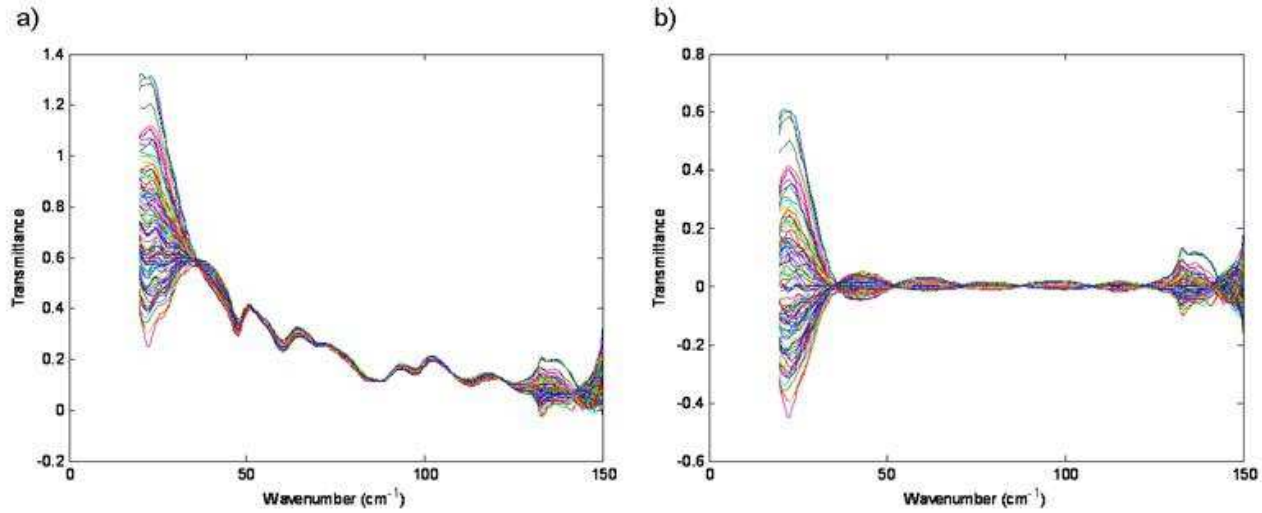

Fig 5: a) T9 block in transmittance in the 20 – 150 cm^{-1} range; b) Column-centered T9 block.

Figure 5 presents all the spectra (outliers excluded) from $\mathbf{T}^9$ block, which is representative of the global behavior. Ideally, the repeatability error should be the same on the whole spectral range. It can be noticed that it is not the case. Firstly, on Figure 5 a), very intense error is spotted at the extremities of the spectra (lower than 30 cm^{-1} and higher than 130 cm^{-1}). This is to be bound to the very low SNR observed Figure 4 b). Secondly Figure 5 b) shows the spectra after centering the block by column. It clearly confirms that the 30 – 130 cm^{-1} range is lowly noisy. However, it appears that there exists a structured variability: the repeatability error depends of the wavenumber.

Wavenumbers 35, 53, 71, 88, 108 and 125 cm^{-1} correspond to nodes where dispersion is minimal, inferior to 1.10^{-3} unit, whereas between two nodes the variability is maximal: +/- 0.05 unit in the 35-53 cm^{-1} range, +/- 0.02 unit in the 53-71 cm^{-1} range, and less than +/- 0.01 unit for the other zones. Hypothesis about the origin of these nodes will be discussed at the end of the article. Moreover, the order of spectra is reversed on the two sides of each node: spectra that presented the higher absorbance become the ones with the lower absorbance after a node and *vice versa*. This phenomenon is clearly visible on Figure 5 b). Node locations are easily spotted, while the variability between two nodes is confirmed.

One can also note that variability is not the same between two “bellies”. This figure also confirms that spectra present extremely noisy extremities where variability is very high.

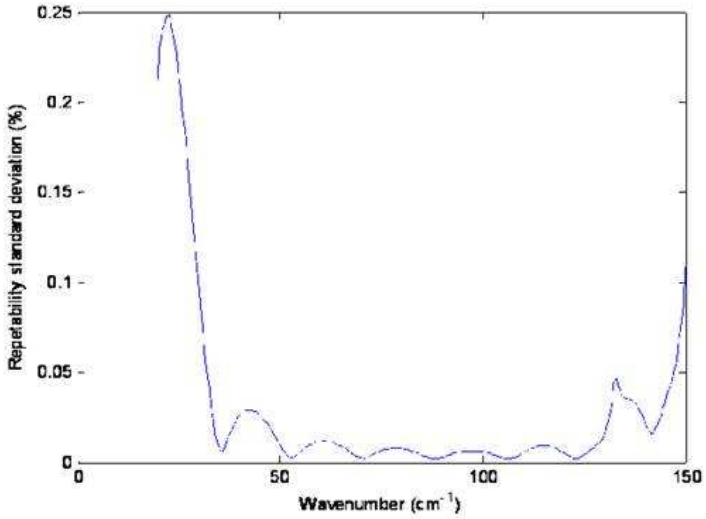

Fig 6: Univariate repeatability error (Within).

Figure 6 shows the evolution of repeatability error as a function of the wavenumber for the complete matrix **T**. As previously, phenomenon are observed: very strong dispersion between 20 and 35 cm^{-1} (nearly 25 times stronger from 0.03 % at 45 cm^{-1} to 0.25 % at 25 cm^{-1}) and above 125 cm^{-1} (from 5 times -0.05 % at 130 cm^{-1}- to 10 times -0.1 % at 150 cm^{-1}- stronger), shown as higher standard deviation compared to the rest of the measure range. Minimum repeatability error has been spotted at 35, 53, 71, 88, 108 and 123 cm^{-1}. They correspond to the nodes identified on Figure 5. Thus, a smaller measure range between 30 and 130 cm^{-1} corresponds to a range where repeatability is better according to the experimental conditions. In addition, each node location spots an inversion point on spectra. All spectra have been impacted by a multiplicative factor that oscillates from a positive to a negative value throughout the measure range, which is null for nodes.

Sensitivity

Device sensitivity is estimated by the relation between the absorption of mean spectra and

the volumetric concentration of a sugar. Figure 7 shows the device sensitivity, calculated as the slope between absorbance and sugar concentration for each wavenumber, for sucrose (a) and glucose (b).

The maximum slope observed for glucose is 0.48 Absorbance Units (A.U.) / $g.L^{-1}$ while the maximum slope for sucrose is 1.3 A.U. / $g.L^{-1}$, which is in contradiction with the spectra furnished by Riken database that shows similar unitary absorbance. This result can be explained by phenomena related to light diffusion. As the granulometry range of our samples (a few hundred microns) is approximately the same size than the wavelength of the THz ray, Mie's law[18] applies. When granulometry increases, so does forward diffusion. Diffusion augmentation implies that photons could encounter more particles and then be more absorbed. Then, as sucrose granulometry is higher than glucose granulometry, the more sucrose, the more diffusion and so the more apparent absorbance because of the optical pathlength increasing.

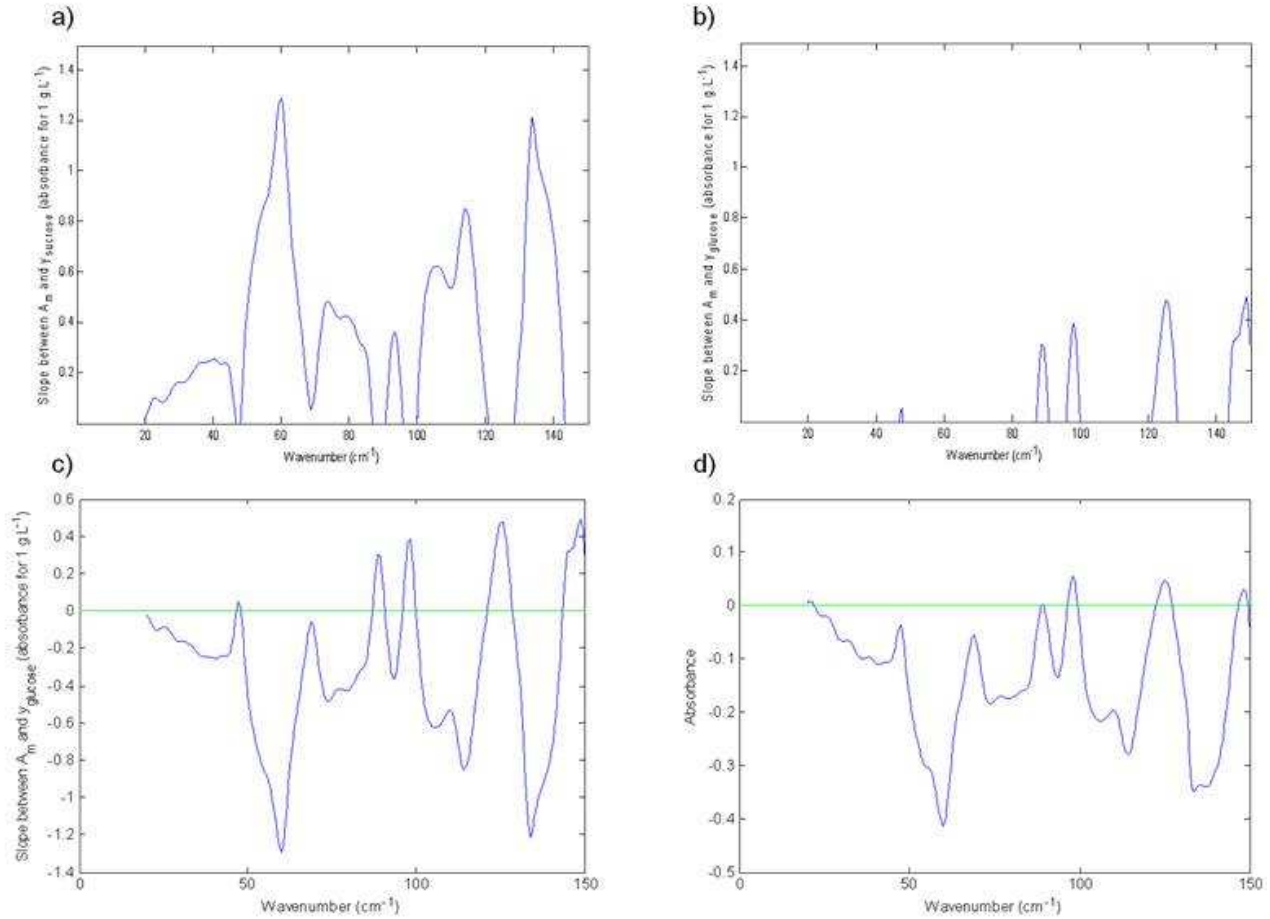

Fig 7: Slope between absorbance and sugar concentration for a) sucrose only; b) glucose only; c) whole slope looked as a function of glucose concentration; d) spectrum obtained by subtracting pure sucrose spectrum to pure glucose spectrum.

Figure 7 a) presents several peaks at 60, 93, 105, 114 and 134 cm^{-1}, for 1.3, 0.36, 0.62, 0.85 and 1.2 A.U. respectively, as well as a zone between 73 and 80 cm^{-1} for 0.45 A.U. approximately. These peaks mean that there is a relation between the volumetric concentration in sucrose and mean absorbance. Peaks at 60, 93 and 114 cm^{-1} correspond to three features of sucrose. The peak at 105 cm^{-1} corresponds to the bottom of the sucrose peak at 113.6 cm^{-1}, which can explain its high variability; moreover, it is cumulated to the sucrose peak located at 112 cm^{-1}. The peak at 134 cm^{-1} is in a low repeatability zone and does not correspond to a sucrose characteristic feature, thus it will then not be taken in account for this study. Figure 7 b) shows the slope between absorbance and glucose concentration in discs for each wavenumber. Several peaks appear at 48, 88.5, 99, 125

and 149 cm^{-1}, for 0.05, 0.3, 0.38, 0.48 and 0.49 A.U. / $g.L^{-1}$, respectively. The first four peaks correspond to peaks identified as glucose, whereas the last peak at 149 cm^{-1} located in a low repeatability zone is discarded. It should also be noted that the whole slope calculated as a function of the glucose concentration (Figure 7 c) bears many resemblances with the spectrum obtained when subtracting the pure sucrose spectrum to the pure glucose spectrum shown in Figure 7 d). It appears that the spectrometer is sensitive enough to detect the differences between the two sugars. The strongest slope for sucrose is observed at 60.6 cm^{-1}, the one for glucose at 125 cm^{-1}. For these wavenumbers, Figure 8 shows the absorbance of each spectrum of matrix **A** as a function of the sugar concentration.

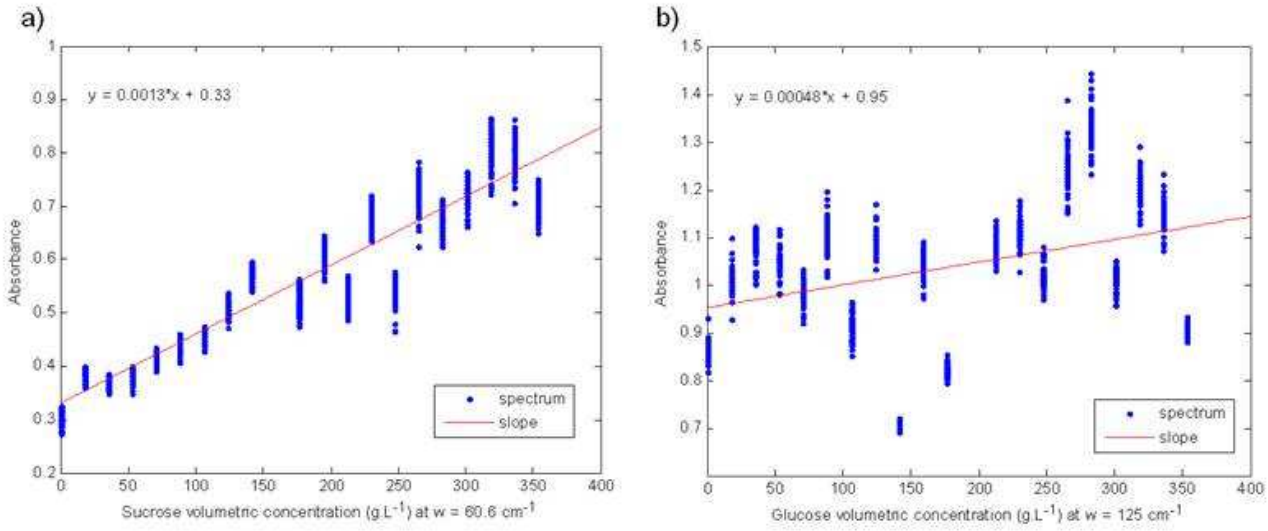

Fig 8: Spectra absorbance for each block at the maximum of sensitivity for a) sucrose; b) glucose.

Figure 8 a) shows a global linear tendency with a slope of 1.3 10^{-3} A. U. per $g.L^{-1}$. In the range of 0 to 124 $g.L^{-1}$ of sucrose, measured absorbance for each block is located near the regression line calculated for 60 cm^{-1}. Between 141 – 230 $g.L^{-1}$, blocks deviate slightly from the regression line, but are still ordered according to the increasing absorbance. It occurs that the blocks with 177 $g.L^{-1}$ and 212 $g.L^{-1}$ of sucrose present similar absorbance. The block $\mathbf{A}^7$,containing 247 $g.L^{-1}$ of sucrose, presents a lower absorbance, with almost 10 %

less than expected. In the range of 265 to 336 $g.L^{-1}$, absorbance is again near the regression line, while the block corresponding to pure sucrose ($\mathbf{A}^1$ with 354 $g.L^{-1}$) is slightly inferior to the line. Besides, variability in absorbance intensity changes from a block to another. For low concentrations in sucrose (0 to 106 $g.L^{-1}$), the gap between the lowest absorbance spectrum and the highest absorbance spectrum in a block is less than 5 %. This gap rises constantly up to 10 % for the 177 $g.L^{-1}$ sample and the four next blocks. Blocks with the highest mass of sucrose (between 265 to 354 $g.L^{-1}$) present gaps around 15 to 20 % differences in absorbance. This is because the peak at 60 cm^{-1} is between two nodes, in a belly where repeatability error is higher, as shown in Figure 6. Furthermore, the variability in absorbance inside a block increases significantly as a function of the sucrose mass present in the mixture. For the highest-containing sucrose samples measures, absorbance present the highest variability, mostly due to a stochastic behavior of the THz ray inside the sample caused by diffusion. It has been checked that there is no relation between absorbance and the order of acquisition and that there is no time-induced shift.

On the contrary, Figure 8 b) shows that sensitivity for glucose studied at 125 cm^{-1} is lower. Some blocks, particularly those containing 141, 177 and 354 $g.L^{-1}$ of glucose, are far away from the regression line, with a difference of approximately 20 % in absorbance. Contrary to figure a), it is difficult to see a clear evolution of repeatability error from block to block as a function of glucose mass. The gap between the highest and the lowest absorbance spectrum goes from 10 to 17 % A.U., with the exceptions of $\mathbf{A}^{16}$ which presents a 21 % gap, and $\mathbf{A}^9$ and $\mathbf{A}^{11}$ which show 3 and 7 % gaps respectively. Sensitivity maximum at 125 cm^{-1} is only 2 cm^{-1} away from a node, hence standard deviation is minimal; this explains why diffusion is almost the same from a block to another. However, Figure 6 illustrates that the repeatability error in this area is increasing quickly; this explains the variability of absorbance between spectra inside a block. We also noted that the glucose

absorbance at 125 cm^{-1} is above 1. This is due to the baseline we observed on spectra which influences the apparent absorbance value.

If an univariate analysis of the measurements is wanted, wavenumber selection is needed by using the relation between mean spectra and sugar concentration. This allows spotting areas where the device is more sensitive towards the analyzed products and where it permits a better prediction. However, device sensitivity is not the same for all the products involved. For our study, it appears that if sucrose is a good candidate to a univariate analysis, it is not the case of glucose. Another method is needed to accentuate information linked to glucose in order to exploit them precisely. Eventually, glucose mass could be deduced from the sucrose mass detected in the samples but if a sample which does not respect the 30 % (m/m) sugar requirement is introduced in a study, univariate analysis will not be able to make the link between the concentration of both sugars. We would like to underscore that we made no attempt in quantification for univariate analysis. Figure 8 might be seen as a first step towards calibration, but we have not defined calibration or test sets as it is done in chemometrics; we used this figure only to represent the lack of precision of univariate methods in comparison with multivariate approaches as we will see in the next part. No quantification attempt will be either discussed in the rest of the article, its feasibility will be reported in a future paper.

Multivariate analysis

The main drawback of a univariate analysis is that it ignores that spectral variables are correlated. Multivariate analysis will bring additive information such as analysis of noise, presence of structured effects, etc.

Signal to Noise Ratio

Calculation of the multivariate SNR on the whole spectral range 20 – 150 cm^{-1} gives a result of 5.41. This means that when considering the global nature of the spectra, signal is only five times superior to noise. This mediocre ratio is the consequence of the extremely noisy extremities of the spectra that are included in the calculation. Applying pretreatments to spectra is a way to improve this ratio. Wavenumber selection, to exclude noisy areas from the calculation while keeping a maximum of zones containing information, suffices to increase significantly the result. Selecting only the variables comprised in the range 50 – 130 cm^{-1}, SNR calculation gives a result of 67.85. This ratio calculated for a smaller range is twelve times superior to the ratio obtained for the whole initial range. As a conclusion, variable selection will be one of the pretreatments to apply to spectra in future studies to optimize the analysis.

Repeatability

Spectra repeatability is studied through the within scatter matrix (**W**) analysis, by the exploitation of **W** eigenvectors and eigenvalues. The first three variance percentages of **W** evolves very quickly: the first one is at 83 %, while the second and the third are only at 3 % and 0.4 % respectively. The following percentages are inferior to 0.1 %. This evolution is characteristic of global systematic effects, such as baseline effects for instance. It also indicates that noise is structured. Most spectral variations due to lack of repeatability are contained in a quasi monodimensional space. This space is supported by the first loading of **W**.

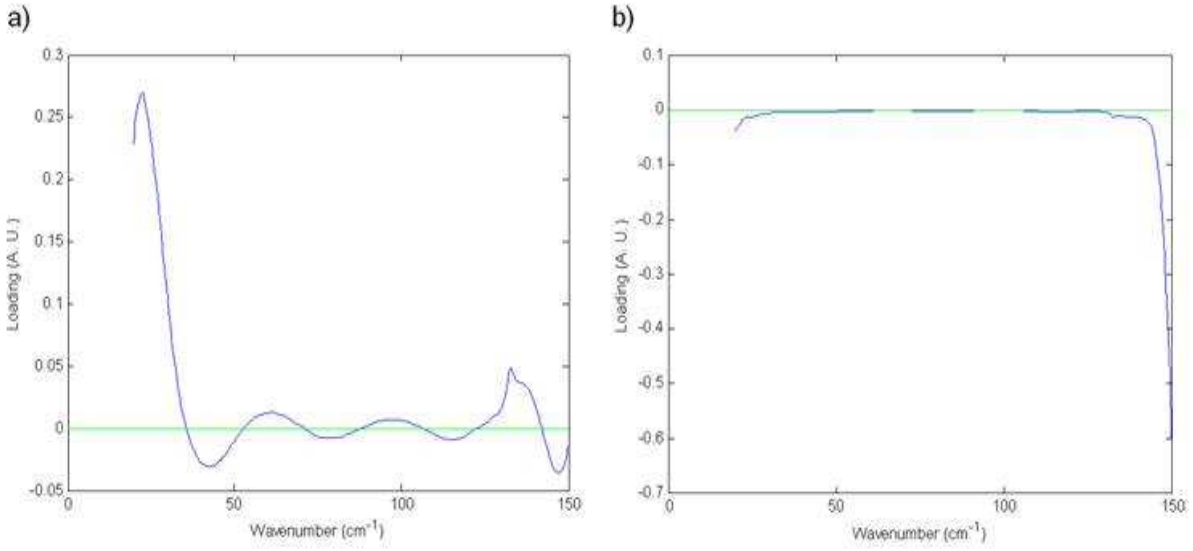

Fig 9: Within scatter matrix W analysis: a) 1st loading, b) 2nd loading.

Figure 9 a) presents the first loading of **W**. High values are observed at the extremities of the loading, i.e. 20 – 30 cm^{-1} and to a lesser extent beyond 130 cm^{-1}, which denotes the strong dispersion of spectra in these area we have already noted. Between 30 and 130 cm^{-1}, a structure in a form of a sinusoid appears. This sinusoidal form confirms the observations of the standard deviation curve seen while studying univariate repeatability (Figure 6). This sinusoid intersects the abscissa axis at the wavenumbers corresponding to the node locations on spectra: 35, 53, 71, 88, 108 and 123 cm^{-1}. Moreover, the sinusoidal form explains the observations made on Figure 5 b) concerning the inversion of spectra after going through a node.

In order to study the stability of this phenomenon, eigenvectors from the variance matrices of all blocks have been calculated and analyzed. Figure 10 shows the superposition of each block's first loading for the whole spectral range (a) and the detail in the range 50 – 130 cm^{-1} (b).

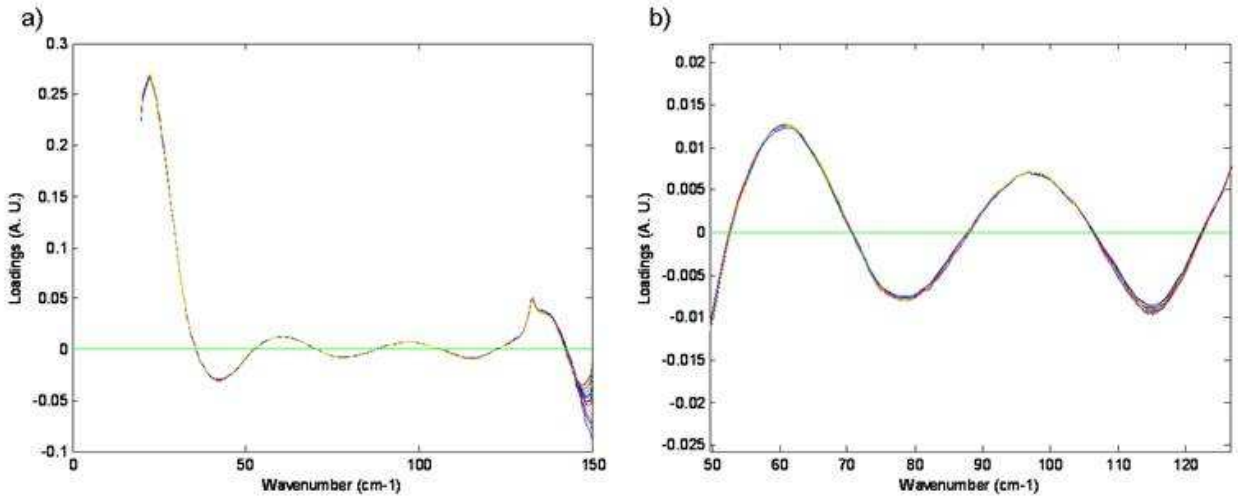

Fig 10: 1st loadings calculated for each block Tk. a) On the whole spectral range, b) on the 50 – 130 cm^{-1} range.

One can see that all loadings are similar in form, corresponding perfectly to the one calculated with all blocks. The sinusoid observed between 50 and 130 cm^{-1} is highly repeatable from one block to another. This demonstrates that noise is structured and reproducible. This phenomenon probably finds its origin in the transformation of the measured interferogram in spectrum during the inverse Fourier transform. The most probable hypothesis would be that a slight interferogram shift would cause a transfer of the continuous component on a low frequency.

W second loading, shown in Figure 9 b), presents a line close to 0 in the range 30 – 140 cm^{-1} meaning that apart from the sinusoidal noise observed on the first loading, the error of repeatability is minimum in this range. Between 20 and 30 cm^{-1}, loading weight (only -0.05) is nonetheless fifty times higher than in the previous range. Beyond 140 cm^{-1}, loading weight decreases very quickly to reach -0.6. These areas correspond to the beamsplitter operational limits (Figure 4 a). Considering the variations observed in a block (Figure 5), we deduced that the second component concurs to the explanation of the variations around the extremities of the spectral range. Noise, as a whole, is supported by the first two loadings of within scatter matrix. The strong repeatability error observed between 20

and 30 cm^{-1} is supported by the first loading while the error observed from 140 to 150 cm^{-1} is supported by the second loading, and thus is independent from the error of repeatability existing in the 20 – 140 cm^{-1} spectral range.

Multivariate analysis demonstrated firstly that strong noise is present at the extremities of the spectra, certainly because of the beamsplitter limits; and secondly is the revelation that the effect applied to spectra in the most repeatable range is a sinusoid. This sinusoid could probably find its origin in the transformation of the interferogram in a spectrum by the inverse Fourier-transform operation or in a default of the mirror mechanism, causing errors due uncontrolled or extensive mirror movements.

Sensitivity

Multivariate sensibility study is realized thanks to the between scatter matrix (**B**) analysis. Variance percentages have been calculated for the first three principal components after matrix diagonalization. As repeatability, they evolve quickly. First percentage is calculated at 74 %, second percentage at 22 % and third percentage inferior at 1 %. First two percentages explain more than 95 % of **B** by themselves.

The two first components are represented by the corresponding loadings in Figure 11, respectively figure a) and b).

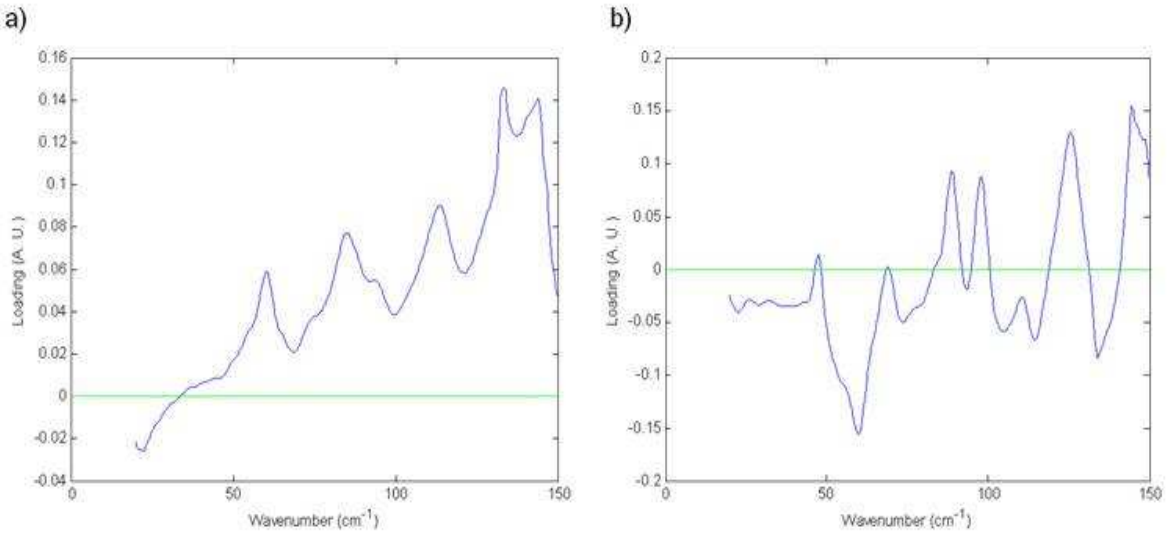

Fig 11: Between scatter matrix B analysis: a) 1st loading, b) 2nd loading.

The first loading is quite similar to the pure sucrose absorbance spectrum (Figure 2 a) while the second loading resembles the univariate sensitivity to glucose (Figure 7 c). This particularity can be explained as follows: in the case of our study, Beer-Lambert's law for a spectrum ***x*** can be written as in Eq. 11.

$$x = C_S \varepsilon_S + C_G \varepsilon_G \quad \text{(Eq. 11)}$$

where C_S and ε_S are respectively the volumetric concentration and the molar extinction coefficient of sucrose, and C_G and ε_G the volumetric concentration and molar extinction coefficient of glucose.

As we are working with pure products, we can assume that a mixture spectrum is a combination between the pure spectra of the products, impacted by their relative concentrations. Therefore, we consider that molar extinction coefficients are assimilated to the corresponding pure sugar spectra. Since $C_G + C_S = 1$, Eq. 11 can be written as Eq.12:

$$x = \varepsilon_S + (\varepsilon_G - \varepsilon_S) C_G \quad \text{(Eq. 12)}$$

As ε_S is close to ε_G, there is a quasi-orthogonality between ε_S and $\varepsilon_G - \varepsilon_S$ which allows

the construction of a two-dimensional plane. We saw that the first loading is represented by the sucrose spectrum, which is the first term of Eq. 12. By construction, the second loading is the difference between glucose spectrum and sucrose spectrum. As it has already been seen, it also corresponds to the spectrum calculated for univariate sensitivity to glucose (Figure 7 c). However, while univariate study showed that sucrose had higher coefficients than glucose, the second loading observed thanks to the multivariate analysis reveals that the maxima for both sugars are equivalent, around 0.15 A. U.. This could mean that while the univariate analysis suggested that the spectrometer was not very sensitive to glucose variations, the multivariate analysis indicates that the spectrometer detects glucose variations as well as sucrose variations from one sample to another.

As it has been mentioned previously in the univariate sensitivity part, we believe that, according to Mie's law, the diameter of sucrose grains has a direct effect on absorption, being more easily detected and increasing the relative absorbance of this product. The second loading indicates that the spectrum identified as univariate sensitivity is no more the main source of variability but the second one. The importance of the differences due to sucrose and glucose variations is minimized compared to the presence of sucrose in the sample.

Net Analyte Signal

The Net Analyte Signal (NAS)[19] is an "ideal" spectrum which varies linearly as a function of the concentration of studied compounds. It represents the net contribution of a product to the signal. One way to obtain the NASs of the two sugars is to unmix the matrix containing the spectra regarding the matrix containing the concentrations.[20, 21]

NAS are preferred to pure spectra because they provide better models. They represent the orientation in the spectral space where the signal varies the more as a function of the

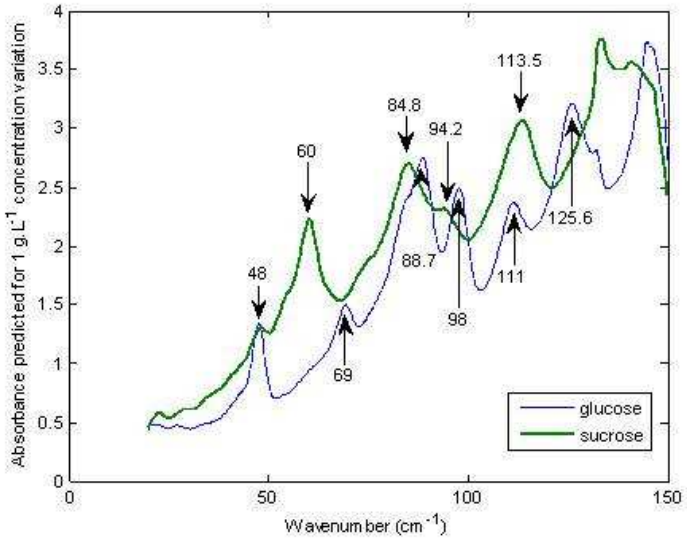

Fig 12: Glucose and sucrose Net Analyte Signals.

concentration of one product. If there is any interaction between products, NAS should take them in account, while pure products spectra could not.

Figure 12 illustrates the reconstituted NAS. Global features from the pure spectra are respected, which denotes an absence of interference or interaction between the two products in mixture. Characteristic peaks of each sugar appear at the expected wavenumbers. NAS utilization shows that the spectrometer sensitivity is good enough to detect the variations of both sugars and link them to spectral variations, enough to recreate the ideal spectra. It means that a multivariate analysis taking in account mixture spectra as well as samples concentrations is able to find useful information relative to measured products. Evaluation of spectrometer sensitivity by NAS consists in comparing the norms of NAS for each product. According to Faber,[17] the more the norm of a NAS, the more important will be its influence on sensitivity. NAS calculations for glucose and sucrose gave the following results:

$NAS_{Glucose}$ norm = 10.84 A.U. / $g.L^{-1}$

$NAS_{Sucrose}$ norm = 12.38 A.U. / $g.L^{-1}$

Sucrose NAS norm is just a little higher than glucose NAS norm. While the univariate sensitivity analysis showed a two to three times higher sensitivity to sucrose compared to glucose, this analysis shows that NAS have similar norms. This means that, according to this multivariate analysis, sucrose does not completely hide the glucose as it was previously suggested by the univariate analysis, but that the spectrometer has equivalent sensitivity for both products. NAS-based multivariate analysis contradicts univariate analysis: variations of both sugars are well detected by the device, independently from the other as the unmixing process resulted in two spectra of similar intensities. It appears that univariate analysis presents serious gaps in apprehending the relation between sugars and spectrometer sensitivity. Figures 7 and 8 suggested that only sucrose variations were clearly seen by the spectrometer, contrary to glucose variations. From this point of view, only sucrose-based models could be constructed, with the condition to work at the maximum of sensitivity at 60 cm^{-1}. The first multivariate study showed that if sucrose is more absorbent, glucose was included in the construction of the loadings of the matrix **B**, which means that glucose variations were indeed detected. Further multivariate analysis with NAS demonstrates that glucose sensitivity is in fact almost as important as sucrose sensitivity. Univariate methods alone are insufficient to study both sugars: only sucrose could be studied, and only for one wavenumber, ignoring the rest of spectrum. To build functional models to analyze both sucrose and glucose, multivariate methods are required to study them not as a component of two products in a mixture, but as separate products.

Conclusion

This work was dedicated to study methodological characteristics of THz spectroscopy on

sugar powder measurements. A design of experiments has been made to measure different mixtures of sucrose and glucose, with repetitions. Metrological aspects of the device have been characterized through the study of SNR, repeatability of measure, and sensitivity of the spectrometer. Results obtained with univariate and multivariate methods have been compared in order to maximize the comprehension of these aspects, which revealed several points.

Firstly, the initial measurement range possesses very noisy extremities. In the range 20 – 50 cm^{-1} and above 130 cm^{-1}, SNR is very low. The error of repeatability is at its maximum in these areas and the source of unreliable data. Univariate and multivariate analysis have shown that the selection of variables should be a solution to improve SNR. The 50 – 130 cm^{-1} range maintains the majority of the spectral characteristic features for both sugars and presents a lower error of repeatability. Secondly, the multivariate analysis of the 50 – 130 cm^{-1} has revealed the presence of a structured noise. This noise, which has the form of a sinusoid, contributes to the error of repeatability. This induces the following question: is it wise to automatically compute the mean spectra if the collected scans are impacted by such noise? For characterization studies, it does not matter as there is no peak shift; but for calibration studies, the variability of spectra could be a source of imprecision. If manual intervention on the device components (i. e. mirrors) does not correct this feature, chemometrics pretreatments should be used to negate or at least attenuate such effects.

For example, Roger *et al.*[22] propose to realize a projection, orthogonal to identified noise (EPO) in order to completely remove the perturbation while conserving useful data. Finally, the results obtained for the sensitivity of the spectrometer are not identical depending on the use of univariate or multivariate methods. univariate analysis suggested that only sucrose was clearly detected by the spectrometer, while the sensitivity to glucose was poor. Meanwhile, the multivariate analysis, and more particularly the reconstitution of NAS,

revealed that the apparatus was equally sensitive to both sugars. As multivariate techniques consider the entire spectra because the variables are correlated, they extract information from every feature, at every variable, and consider it as a whole. As a matter of fact, they do not rely on a single variable and its associated data. Where univariate analysis is limited by the sharpness of the spectra, multivariate analysis can ease the interpretation of sensible zones. We thus recommend using multivariate techniques and pretreatments over univariate methods for an accurate spectral analysis. Further work will address the quantification of powder mixtures based on such multivariate techniques. Various pretreatments have still to be studied to evaluate what benefits they can add to the calibration study.

References

1. M. F. Kimmit, Journal of Biological Physics **29,** 2, 77 (2003).

2. B. M. Fischer, M. Walther and P. Uhd Jepsen, Phys. Med. Biol. **47,** 21, 3807 (2002).

3. D. F. Plusquellic, K. Siegrist, E. J. Heilweil and O. Esenturk, Chem. Phys. Chem. **8,** 17, 2412 (2007).

4. P. F. Taday, Phil. Trans. R. Soc. Lond. A **362**, 351 (2004).

5. C. J. Strachan, P. F. Taday, D. A. Newnham, K. C. Gordon, J. A. Zeitler, M. Pepper and T. Rades, J. Pharm. Sci. **94,** 4, 837 (2005).

6. A. G. Davies, A. D. Burnett, W. Fan, E. H. Linfield and J. E. Cunningham, Materials Today **11,** 3, 18 (2008).

7. J. F. Federici, B. Schulkin, F. Huang, D. Gary, R. Barat, F. Oliveira and D. Zimdars, Semiconductor Science and Technology **20,** S266 (2005).

8. J. Hooper, E. Mitchell, C. Konek and J. Wilkinson, Chemical Physics Letters **467,** 4-6, 309 (2009).

9. D. G. Allis, J. A. Zeitler, P. F. Taday and T. M. Korter, Chemical Physics Letters **463,** 1-3, 84 (2008).

10. G. F. Liu, X. J. Ma, S. H. Ma, H. W. Zhao, M. W. Ma, M. Ge and W. F. Wang, Chinese Journal of Chemistry **26,** 7, 1257 (2008).

11. Y. Hu, P. Huang, L. Guo, X. Wang and X. C. Zhang, Physics Letters A **359,** 728 (2006).

12. Y. Zhang, X. H. Peng, Y. Chen, J. Chen, A. Curioni, W. Andreoni, S. K. Nayak and X. C. Zhang, Chemical Physics Letters **452,** 1-3, 59 (2008).

13. M. Ge, H. Hongwei, T. Ji, X. Yu, W. Wang and W. Li, Sci. China Ser. B **49,** 3, 204 (2006).

14. L. Yang, H. Sun, S. Weng, K. Zhao, L. Zhang, G. Zhao, Y. Wang, Y. Xu, X. Lu, C. Zhang, J. Wu and C. Jia'er, Spectrochimica Acta Part A **69,** 160 (2008).

15. H. Martens, T. Naes, *Multivariate Calibration* (John Wiley & Sons Ltd, Chichester, 1989).

16. D. Bertrand, E. Dufour, *La spectroscopie infrarouge et ses applications analytiques*

Tec & Doc, Ed. (Lavoisier, Paris, 2000), 2nd ed., p. 7.

17. N. K. Faber, Anal. Chem. **71,** 3, 557 (1999).

18. G. Mie, speziell kolloidaler Metallösungen. Ann. Phys. Leipzig **25,** 377 (1908).

19. A. Lorber, K. Faber and B. R. Kowalski, Anal. Chem. **69**, 8, 1620 (1997).

20. S. M. Short, R. P. Cogdill and C. A. Anderson, AAPS PharmSciTech **8,** 4, Article 96 (2007).

21. R. Bro and C. M. Andersen, J. Chemom. **17,** 646 (2003).

22. J. M. Roger, F. Chauchard and V. Bellon-Maurel, Chemometrics and intelligent laboratory systems **66,** 2, 191 (2003).

3ème publication : Étalonnage en spectroscopie Terahertz appliquée à des contenus de sucres

Introduction

Comme il a été dit précédemment, la caractérisation de produits par spectroscopie THz est simplifiée car il n'y a pas d'information superflue provenant de la matrice d'échantillonnage. La présence d'impuretés, des modifications structurelles dues à des changements de phases ou à des conditions différentes lors de la préparation des échantillons peuvent provoquer une modification du spectre du produit, ce qui pourrait rendre la spectroscopie THz précieuse dans une optique de suivi de réaction, de contrôle-qualité ou encore de détection de contrefaçon.

Plusieurs publications, listées dans l'article suivant, se sont concentrées sur l'acquisition de données, tentant de connecter les pics apparaissant sur les spectres aux groupes chimiques présents dans les molécules ou de prédire où ces bandes devraient apparaître en utilisant des programmes de prédiction développés pour des méthodes infrarouges. Ces travaux ont réuni une quantité précieuse de données spectrales mais ont obtenu des résultats aux succès relatifs en ce qui concerne l'aspect prédiction. Il s'avère que les vibrations intermoléculaires caractéristiques des fréquences THz ne peuvent pas être prédites de la même manière que les vibrations intramoléculaires, du moins à ce jour.

Faute de pouvoir prédire efficacement les positions des bandes caractéristiques des molécules, il est alors nécessaire de constituer une base de données spectrales avant d'envisager la quantification et la création de modèles de prédiction pour quelques produits que ce soit. Le Terahertz étant une technique récente, peu de publications se sont tournées vers cet aspect, et la plupart se sont contentées de modèles simples. Cet article désire mettre en évidence la faisabilité de quantifier des produits sous forme de poudres à l'aide de la spectroscopie THz, à la fois purs et en mélange.

Matériels et Méthodes

Dispositif expérimental

Le spectromètre utilisé était un appareil de type « Bruker IFS 66v/S FTFIR », équipé d'une séparatrice 23 µm en Mylar dont la gamme d'efficacité était comprise entre 150 et 20 cm^{-1} (0.6 - 4.5 THz). La source était une lampe au mercure émettant un rayonnement continu entre 600 et 5 cm^{-1} (0.15 – 18 THz). Le détecteur était un bolomètre au silicium de type « Infrared Laboratories Model N° HD-3 » refroidi à l'Hélium liquide à 4.2 K avec un filtre passe-haut, couvrant la gamme 0.5 - 10 THz. Le diaphragme présentait une ouverture de 12.0 mm. La vitesse de balayage était fixée à 4 Hz. La résolution utilisée était de 0.5 cm^{-1}. Toutes les mesures ont eu lieu sous vide et à température ambiante, 296 +/- 1 K.

Une référence a été prise toutes les heures, avec une pastille de PE pure compactée de la même manière que les autres échantillons (voir ci-dessous). Chaque échantillon a ensuite été placé dans le spectromètre, sous vide, de façon à ce que le point de focalisation du rayon se situait au centre de la surface du comprimé faisant face à la source. Chaque spectre correspondait à la moyenne de 100 scans.

Les échantillons étaient constitués de mélanges ternaires de poudres de polyéthylène (PE), saccharose et glucose et de mélanges binaires entre le PE et soit le glucose ou le saccharose séparément. Le saccharose a été obtenu auprès de Fluka Analytical (Ref. N° 84100) et le glucose auprès de Sigma-Aldrich (Ref. N° 16325). Aucun n'a nécessité de purification supplémentaire. Le glucose présentait une granulométrie d'environ 100 µm et a été utilisé tel quel tandis que le saccharose a été broyé finement dans un mortier avant utilisation (granulométrie < 0.5 mm). Le PE a été acquis chez Aldrich (Ref N° 26935-2, spectrophotometric grade powder). La quantité de PE a été fixée à 70 % (m/m) de la masse totale des échantillons. Les 30 % (m/m) restants ont été composés d'un mélange entre saccharose et glucose selon 21 pourcentages massiques : le premier échantillon contenait 0 % de glucose et 100 % de saccharose, le second 5 % de glucose et 95 % de

saccharose, et ainsi de suite par pas de 5 % relatifs, jusqu'à 100 % de glucose. Les mélanges ont été ensuite compactés sous forme de pastilles de 0.5 ± 0.2 mm d'épaisseur et 12 mm de diamètre, pour un volume moyen de 56.55 mm^3. Chaque pastille a ensuite été re-broyée puis re-compactée, afin d'améliorer son homogénéité. La masse de chaque pastille était de 70 ± 0.1 mg.

Des pastilles de sucres purs ont été préparés pour compléter notre set d'étalonnage. Elles étaient composées de poudre de PE et saccharose pur ou de glucose pur. Les masses de glucose utilisées ont été de 1, 5*, 7, 9, 10, 13, 15*, 17 et 20* mg. Celles de saccharose ont été de 1, 3*, 4, 5*, 7, 9*, 11, 13, 15*, 17, 18 et 20* mg. Les échantillons marqués d'une astérisque ont été répétés. Le poids total des pastilles était toujours de 70 ± 0.1 mg au final.

Data Processing

Les spectres moyens ont été collectés en transmission grâce au logiciel *Opus/IR*. Les échantillons purs ont été regroupés dans le set d'étalonnage tandis que les mélanges ont été regroupés dans le set de test. Les outliers dans le set d'étalonnage ont été détectés et exclus lors des ACP réalisées avec le logiciel *The Unscrambler* (CAMO, version 9.2). Les échantillons exclus étaient ceux présentant une forte variance résiduelle ou dépassant la limite T^2 du test de Hotelling. Les modèles ont été construits et testés avec Matlab (version 7.4.0 R2007a). La régression PLS avec validation croisée sur le set d'étalonnage, la régression basée sur la sélection de variables selon la méthode CovSel [Roger 2011] et la régression basée sur les NAS [Faber 1999] ont été testées.

Les écarts-types cités dans l'article (SECV, SEP_C) et le biais ont été calculés de la manière suivante :

$SECV = RMSE(\hat{y} - y)$ (Eq.13);

$SEP_C = RMSE(\hat{y} - y - biais)$ (Eq. 14);

$Biais = Mean(\hat{y} - y)$ (Eq. 15).

Avec : RMSE : Root Mean Square Error, y la valeur réelle en sucre d'un échantillon et ŷ la valeur prédite pour cet échantillon.

Présentation de l'article

Le but de cet article est de déterminer si des méthodes de prétraitement chimiométriques peuvent être appliquées à la spectroscopie THz en vue d'une éventuelle quantification. Nous en profitons pour essayer quelques méthodes de prétraitement peu connues (CovSel, Net Analyte Signal). Ces méthodes sont comparées afin de déterminer si un modèle peut être a priori une solution de choix concernant l'analyse d'un sucre en particulier ou s'il convient aux deux. Un modèle qui accentuera la distinction entre les sucres sera favorisé, étant donné la proximité des bandes communes aux deux produits.

Discussion

La projection des échantillons contenant les mélanges sur l'Analyse en Composantes Principales (ACP) réalisées sur le set de calibration, c'est-à-dire les échantillons composés uniquement de sucre et de polyéthylène, est montrée dans la Figure 11.

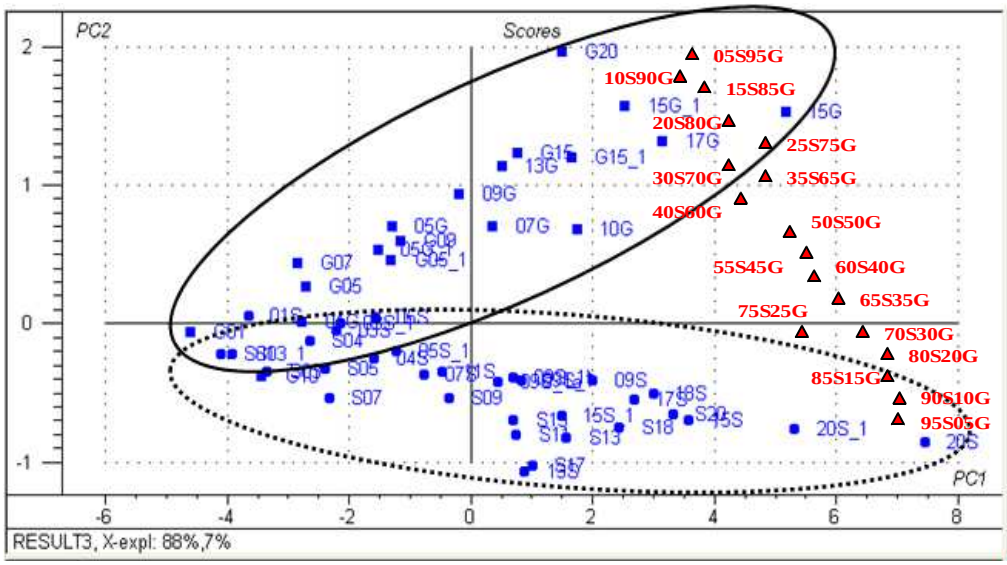

Figure 11: Projection des échantillons de mélanges (en rouge) sur l'ACP des échantillons de sucres seuls (en bleu, glucose dans le cercle plein, saccharose dans le cercle en pointillés).

Nous constatons que l'ensemble de nos échantillons forme un triangle de mélange, c'est-à-dire que les échantillons de sucres seuls forment deux arêtes du triangle, chacune portée par une combinaison des deux premières composantes principales, et que les mélanges forment le troisième

côté, reliant les sommets constitués des échantillons contenant 20 mg de sucre au total. Il y a bien un effet de masse influençant l'ACP. Nous pouvons donc estimer que la quantification des sucres sera possible à ce stade de l'étude.

Deux choix s'offrent à nous :

- Composer un set de calibration comprenant à la fois des échantillons de sucres seuls et des mélanges, et un set de test sur le même principe. Cette approche nous permettrait d'étudier la capacité de distinction des sucres de la méthode vis-à-vis des mélanges et voir si elle peut quantifier des petites concentrations en sucres.
- Composer le set de calibration constitué des sucres seuls uniquement, et le set de test constitué des mélanges. Cela permet d'inclure les échantillons de faible concentration dans la calibration, afin de déterminer si le modèle est capable de les prédire avec peu d'erreur. D'autre part, cela représente ce que serait une étude de cas pratique d'analyse de résidus de pesticides sur des aliments : à partir d'une base de données comprenant généralement des molécules pures, il faudra que la méthode de quantification retenue soit capable de différencier les molécules et de les quantifier, quelle qu'en soit la proportion.

Cette deuxième approche a été retenue et est présentée dans l'article suivant.

En complément des méthodes testées et présentées dans la publication, nous avons également calculé des modèles de régression basés sur la PLS 2 et sur l'algorithme CovSel appliqué sur les spectres prétraités par une dérivée seconde de Savitzky-Golay.

Les erreurs SECV des modèles de calibration créés par PLS 2 sur les spectres bruts par cross validation sont du même ordre de grandeur que les autres modèles (cf. Tableau 2), avec des SECV égales à 33.8 $g.L^{-1}$ pour le saccharose et 25 $g.L^{-1}$ pour le glucose. Les résultats du modèle prédictif PLS 2 sont montrés dans la Figure 12. Nous constatons que les échantillons de mélanges testés montrent de forts biais pour la prédiction du saccharose comme pour celle du glucose. Néanmoins,

les coefficients de détermination sont proches de 1, ce qui suggère qu'une correction de type biais-pente sera efficace sur ces modèles et améliorera la prédiction.

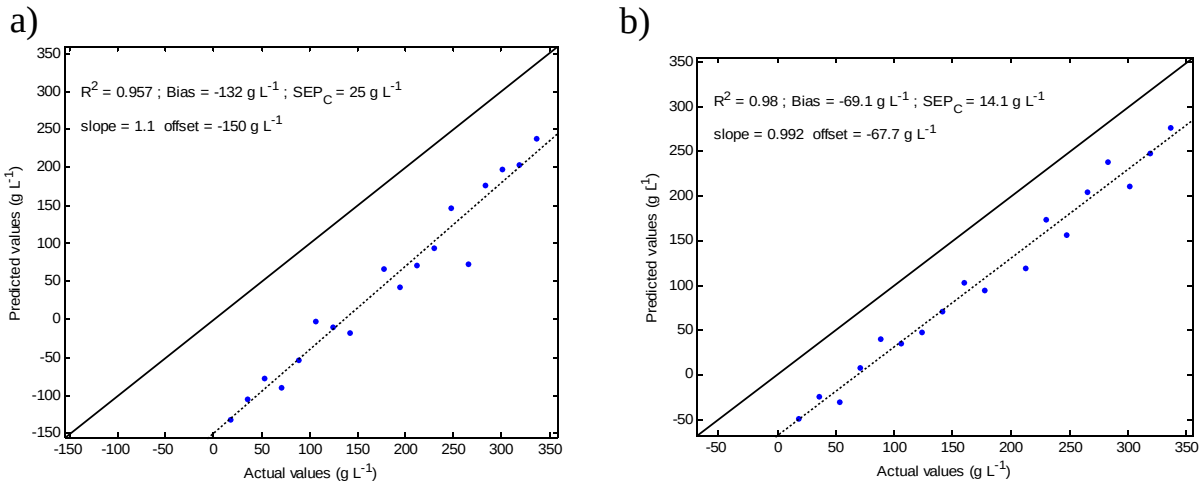

Figure 12 : Tests du modèle PLS 2 sur les échantillons de mélange. a) Concentrations en saccharose ; b) Concentrations en glucose.

Nous constatons également que les erreurs de prédiction obtenues avec ces modèles sont parmi les plus faibles sur l'ensemble des modèles testés. La gestion des réponses du glucose et du saccharose dans un seul traitement chimiométrique a amélioré le modèle prédictif ; nous faisons alors l'hypothèse que bien que les pics caractéristiques de chaque produit soient apparus inchangés (en termes de position de pic) dans les spectres de mélange, l'importance d'une ou plusieurs interactions chimiques qui existeraient entre les sucres ne doit pas être négligée et qu'idéalement les réponses ne doivent pas être traitées séparément. L'amélioration des résultats obtenus par PLS 2 par rapport à la PLS 1 suggérerait que les vibrations intermoléculaires observées sur un spectre THz impliquent une plus grande complexité du système et qu'il est nécessaire de prendre tous les composants de celui-ci en compte pour rendre la prédiction la plus précise possible. L'étape suivante de la calibration, mais qui n'a pas été envisagée ici, serait de tester la PLS 2 sur les spectres impactés par une dérivée seconde : étant donné que ce prétraitement a amélioré les modèles PLS 1, il serait intéressant de déterminer s'il en va de même en PLS 2.

La deuxième méthode que nous souhaitons rajouter à la discussion de l'article est l'utilisation de la méthode de sélection de variables CovSel aux spectres auxquels a été préalablement appliquée une

dérivée seconde de Savitzky-Golay. Dans une même optique qu'en PLS 1, nous avons voulu appliquer ce prétraitement afin de déterminer si la sélection de variables s'en trouve impactée et améliore ou non le modèle. Les résultats sont montrés Figure 13.

En comparant ces résultats aux modèles créés par CovSel sur les spectres bruts (Fig. 7 et Fig. 8 de la publication), nous pouvons constater que la dérivée seconde n'apporte pas d'amélioration ni aux modèles de calibration, ni aux prédictions. La dispersion des échantillons à 0 $g.L^{-1}$ ne contenant pas de saccharose est beaucoup plus importante qu'avec le modèle développé sur les spectres bruts et pour les deux sucres les SECV sont plus grandes. Quant aux tests, les SEP_C sont également plus importantes pour le modèle avec dérivée (10 $g.L^{-1}$ d'erreur en plus) que sans. Nous voyons donc que l'utilisation de la dérivée seconde n'a pas d'influence bénéfique sur les modèles développés avec l'algorithme CovSel, contrairement à la PLS 1.

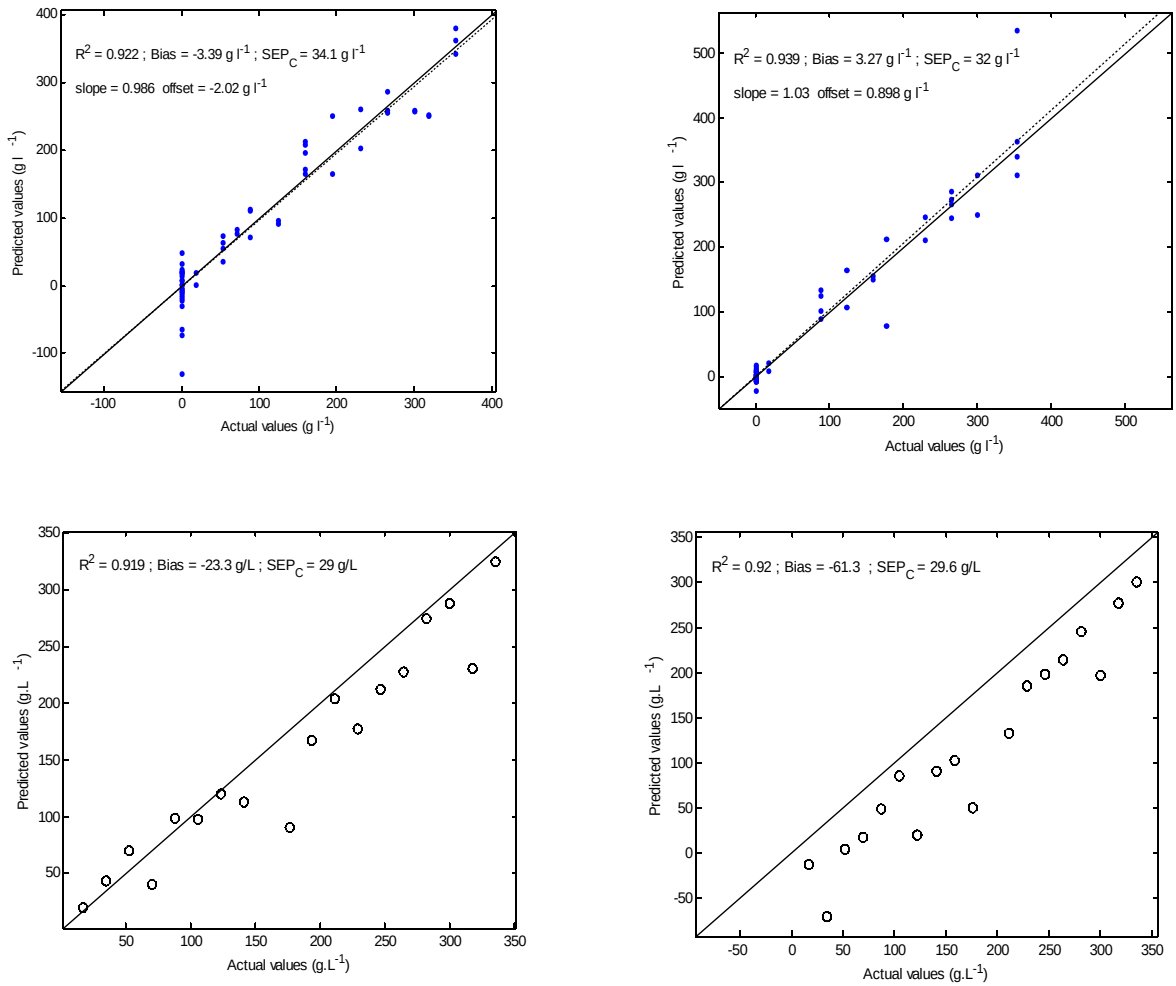

Figure 13: Résultats du modèle développé par la méthode CovSel, les spectres ont été prétraités avec une dérivée seconde. En haut calibration, en bas test ; colonne de gauche : saccharose, colonne de droite : glucose. Note : il y a une erreur dans la notation du programme : il faut remplacer dans les cadres du haut SEP_C par SECV.

Le Tableau 2 regroupe les différents résultats des modèles testés et retenus, en comparant le nombre de variables latentes retenues pour chaque modèle, l'erreur standard de calibration, celle de test et le coefficient de détermination des prédictions.

Méthode	Produit	Variables retenues	Calibration (SECV en g/L)	Test (SEP_C en g/L)	R^2
PLS 1 avec dérivée seconde	Saccharose	8	27,6	45	0,795
	Glucose	7	19,6	24	0,975
NAS	Saccharose	-	41	57,9	0,928
	Glucose	-	19,5	18,1	0,971
CovSel	Saccharose	7	30,7	18,1	0,924
	Glucose	6	21,5	19,5	0,962
CovSel avec dérivée seconde	Saccharose	8	34,1	29	0,919
	Glucose	6	32	29,6	0,920
PLS 2 avec dérivée seconde	Saccharose	5	33,8	25	0,957
	Glucose	5	25	14,1	0,980

Tableau 2: Variables latentes, erreurs standard et coefficients de détermination des modèles en calibration et en prédiction pour chaque sucre en fonction des méthodes de prétraitement.

Le modèle créé par PLS 1 présente les meilleurs modèles de calibration si nous souhaitons prendre en compte le saccharose et le glucose en même temps. Toutefois, la prédiction des échantillons de mélange a montré que l'erreur des modèles PLS 1 est élevée, tout particulièrement pour le saccharose. Ce modèle ne suffit pas pour quantifier correctement les sucres présents dans nos mélanges.

L'utilisation des NAS a conduit à l'élaboration d'un modèle très performant concernant le glucose, mais avec une erreur beaucoup trop importante pour le saccharose. Il pourrait donc être envisageable d'utiliser les NAS dans le but de quantifier le glucose uniquement dans les mélanges, tandis qu'une autre méthode devra être retenue pour le saccharose. Il est intéressant de mettre ce résultat en parallèle avec l'étude de sensibilité qui a été menée auparavant puisqu'il est confirmé ici que l'utilisation d'une technique d'analyse multivariée élaborée telle que la détermination du Net

Analyte Signal a permis de montrer que la sensibilité du spectromètre au glucose n'est pas inférieure à celle au saccharose. Bien au contraire puisque le modèle obtenu pour le glucose est l'un des plus précis de ceux que nous avons testé.

La méthode CovSel a été testée dans deux cas : sur les spectres bruts, et sur les spectres prétraités avec une dérivée seconde de Savitzky-Golay. Ce deuxième cas a été envisagé après avoir constaté l'amélioration des modèles de prédiction PLS 1 après l'utilisation d'une dérivée. Néanmoins, nous constatons que la dérivée n'apporte rien aux modèles qui utilisent l'algorithme CovSel. Les erreurs de calibration augmentent de plus de 3 g/L pour le saccharose et de près de 10 g/L pour le glucose, tandis que les erreurs de prédiction augmentent de 10 g/L pour les deux sucres. Il apparaît clairement qu'un prétraitement des spectres qui a été bénéfique à la méthode de régression PLS ne l'est pas nécessairement pour une autre méthode. En l'occurrence, dans notre cadre expérimental, l'algorithme CovSel donne de meilleurs résultats sur les spectres bruts.

Enfin, étant donné la relation qui existe entre les deux sucres au sein des mélanges, une régression de type PLS 2 a été utilisée. Il s'avère que si les modèles de calibration ont des SECV supérieures aux modèles obtenus par PLS 1 (ce qui est lié au fait que le modèle PLS 2 doit prendre en compte les réponses des deux sucres, et non plus un seul à la fois), la validation des modèles montre que les SEP_C pour les deux sucres sont inférieures à celles de la PLS 1. On aurait pu s'attendre à ce que les SEP_C soient supérieures étant donné justement le fait que la prédiction tient compte des deux sucres simultanément en PLS 2, ce qui est source d'erreur. Ces résultats étant arrivés dans les ultimes jours de la thèse, nous n'avons malheureusement pas pu formuler d'hypothèse qui nous satisfasse pour expliquer ce phénomène. Peut-être est-ce lié aux vibrations intermoléculaires détectées en THz, ce qui impliquerait une plus grande interaction entre les produits qui se répercuterait sur la PLS 2 puisqu'elle gère les deux réponses.

Après ces analyses, il apparaît que la sélection de variables CovSel est globalement la méthode la plus prédictive dans notre étude pour les deux sucres (pour le glucose seul, les NAS sont plus

indiqués). On peut également noter que la PLS 2 fournit des résultats très proches, avec seulement 5 variables latentes contre 7 et 6 (pour le saccharose et le glucose respectivement) pour CovSel. Ceci pourrait être un argument en faveur de la PLS 2 par rapport à CovSel si l'on cherche à éviter d'employer trop de variables pour décrire le système et le sur-modéliser.

Conclusion

Cet article a permis de montrer que les techniques chimiométriques qui sont habituellement utilisées en spectroscopie proche infrarouge peuvent être transposées sans inconvénient à la spectroscopie THz dans le cadre de la mesure des poudres. Nous avons pu déterminer qu'un modèle PLS sur les sucres, créé après que les spectres aient été dérivés, n'est pas le meilleur modèle en ce qui concerne la prédiction de la concentration en sucres dans un mélange ; et ce bien que les biais et les SEP_C soient les plus faibles. De plus, l'ajout de la SNV (qui est parfois appliquée « par défaut » lors des constructions de modèles) a eu l'effet inverse de celui attendu en abaissant la qualité du modèle à cause de la normalisation. Dans notre cas, c'est l'information spectrale provoque elle-même un effet multiplicatif sur les spectres, il ne s'agit donc pas d'un bruit que l'on souhaiterait retirer par une SNV. D'autres méthodes chimiométriques permettent d'obtenir des modèles sensiblement équivalents en étalonnage mais bien meilleurs en termes de quantification. La méthode de sélection de variables CovSel est adaptée pour les deux sucres et donne les meilleurs modèles. Par ailleurs, alors que l'utilisation des NAS fonctionne bien sur les échantillons de glucose, nous avons vu que cette technique peut encore être améliorée pour la quantification de saccharose en appliquant une correction biais-pente. Toutefois, la variation du saccharose pour les échantillons à forte concentration peut être un facteur rédhibitoire. Néanmoins, cette technique est la seule qui a permis de réduire significativement la dispersion des échantillons purs à 0 $g.L^{-1}$ de sucre (pour les deux modèles), donc la discrimination de ces produits est plus efficace. Nous

pouvons également mettre ces résultats en corrélation avec ceux de l'étude précédente, où l'analyse de spectres par le NAS a permis d'établir que la sensibilité du spectromètre au glucose était d'un bon niveau, contrairement à ce que les autres méthodes suggéraient. Il en découle que les NAS pourraient être une piste à suivre pour l'analyse de produits dont les spectres sont proches, facilitant leur distinction et la quantification. Le but de l'article étant de démontrer la faisabilité de la quantification de poudres par spectroscopie THz, l'étude est volontairement restée limitée aux premiers résultats obtenus. Nous n'avons pas voulu optimiser outre mesure nos modèles, ni n'avons exploré toutes les méthodes chimiométriques à notre disposition (Principal Component Regression, Multi Linear Regression...) sur ces systèmes. D’autres études devront être réalisées d'une part pour approfondir ces connaissances, mais aussi porter d'autre part sur des mélanges plus complexes pour confirmer le fait que l'ensemble du panel des méthodes chimiométriques peut être appliqué à la spectroscopie THz et qu'il est possible de conserver une bonne aptitude à la quantification des différents constituants des mélanges.

Publication

as submitted in ***Analytica Chimica Acta***

CALIBRATION OF TERAHERTZ SPECTROSCOPY APPLIED TO SUGAR CONTENTS

M. Papillaud*[a], C. Gergely[b,c], F. Davrieux[a], A. Kapitan[a], J.M. Roger[d]

[a] Centre International pour la Recherche Agronomique et le Développement (CIRAD), UMR 95 Qualisud, TA B-95/16, 73 rue J. F. Breton 34398 Montpellier Cedex 5, France

[b] Université Montpellier 2, Laboratoire Charles Coulomb UMR 5221, F-34095, Montpellier, France

[c] CNRS, Laboratoire Charles Coulomb UMR 5221, F-34095 Montpellier, France

[d] UMR ITAP, Cemagref, BP 5095, Montpellier Cedex 1, France

*Corresponding author: Matthieu Papillaud, CIRAD UMR Qualisud, TA B-95/16, 73 rue J. F. Breton 34398 Montpellier Cedex 5, France. Phone: +33(0)4 67 61 54 52; Fax: +33(0)4 67 61 44 33. Email: mpapillaud@yahoo.fr

Abstract

Terahertz spectroscopy is a recent and continuously developing technology which has proven to be a new asset in the analytical range. During the last years, the majority of papers have focused on the identification of different molecules. Only a few have dealt with the calibration of the products they study. The aim of this paper is to create quantification models of sugar concentration in mixtures and to determine whether chemometrics can be applied to Terahertz spectroscopy as it is used in other analytical spectroscopic techniques. Along with Partial Least Square regression, we present calibration techniques which may not be well-known, despite their effectiveness: CovSel variable selection and Net Analyte Signal methods. The CovSel technique, while providing slightly inferior calibration models, shows the better prediction for sugar concentration in mixture. On the other hand, the Net Analyte Signal method creates better prediction models for samples containing only one sugar, while the prediction of mixtures needs further adjustments to improve the models.

Keywords

Chemometrics; partial least square (PLS) regression; variable selection; CovSel; Net Analyte Signal (NAS); Terahertz (THz) spectroscopy.

1- Introduction

Terahertz (1 THz = 10^{12} Hz) technology is a continuously developing technique since the 1980s. Constant innovation in the semi-conductors area guaranteed the success of this technique through the development of brighter sources and detectors that are more sensitive to THz waves and user-friendly. Many domains, including chemistry, biology and pharmaceutical industry, are using THz radiation to fill the gaps in various spectroscopic domains or complement other known and well-used techniques such as infrared spectroscopy.

THz frequency domain covers approximately the region from 100 GHz to 30 THz, i.e. the wavenumbers between 3 and 1000 cm^{-1}. Many materials have a unique spectral fingerprint in this region whereas some others can be transparent or semi-transparent (textiles, paper, wood, cardboard, plastics…). THz rays are weakly energetic and non-ionizing which renders them harmless. The strong absorption of THz waves by liquid water suggests a good interaction between biological samples and THz waves. These waves induce rotation and vibration of polar water molecules and excite low energy intermolecular bonds, such as hydrogen bonds, in water and proteins for example.

Characterization is simplified as there is no superfluous information due to the nature of the matrix that contains the sample, but it has to be mentioned that THz can be sensitive to the preparation method if structural changes happen or impurities are introduced. The chemical sensitivity of the THz technique has contributed to develop security applications: recognition of different products put in sealed containers without direct contact [1] or explosives detection [2-5]. Several publications have focused on data acquisition, trying to relate the different peaks appearing in the spectra to the chemical groups present in the molecule or to predict where the bands should appear by using prediction programs

developed for infrared methods. These works have collected valuable data for chemical recognition but have had more diverse results when it came for spectra prediction or interpretation [6].

A major stake in chemical industry, isomer differentiation can be achieved quite easily and require no further investigation [7]. Therefore, different molecules with even the lightest difference in their molecular structures, which would not be discerned easily with other analysis methods, will get different THz spectra. In this study, Taday has shown that the three molecules analyzed, of very similar in structure, present strong features that are not common to every spectrum because of the low vibrational mode which concerns the entire molecule, and not only groups of it. THz spectroscopy is also sensitive to minor changes in molecular conformation. Epimers (diastereomers that differ in configuration of a single asymmetric carbon) have also shown different spectra when characterized by THz [8]. Walther *et al.* have worked in a similar way than chemists, characterizing the non-covalent intermolecular forces in sugars in polycrystalline and amorphous states. The results of this study were satisfying enough, THz enabling the distinction of the different forms quite easily [9]. These studies prove that THz spectroscopy is a very sensitive method for the differentiation of similar products.

THz frequencies are concerning intra- and inter-molecular vibrations, as well as hydrogen bonding. Therefore, any difference in materials processing history, i.e. changes in the process that could induce a structural or a conformational change, can be detected by this method and will result in changes among the spectrum. This makes THz ideal as a Process Analytical Technology as well as IR and NIR in chemical and especially pharmaceutical industries [7]. Many works have transposed IR principles, combining THz and chemometrics pretreatments to evaluate and quantify the proportions between active ingredients and excipients, for example [10]. The levels of polymorphism and crystallinity of

drugs have also been studied, quantifying each form during real-time analysis in order to optimize the fabrication process [11]. Reactions are monitored and can be adapted to the situation: diastereomers can be identified and quantified [8], which means that a reaction can be stopped at the moment when all conversions into one form has been done, thus being less time- or product-consuming.

To our knowledge, no other paper has presented any work about the quantification of sugars in a mixture. It is but one of the first steps towards the increasing complexity of the systems studied with THz spectroscopy. Also, no other study has been focused on the utilisation of chemometrics techniques to create prediction models. Only Partial Least Square regression models have been presented in some of the papers we mentioned earlier. We would like to present other methods and their use in THz spectroscopy.

2- Notation

Capital bold characters will be used for matrices, e. g. **X**; small bold characters for column vectors, e. g. $\mathbf{x}_i$ will denote the i^{th} column of **X**; row vectors will be denoted by the transpose notation, e. g. $\mathbf{x}^T_j$ will denote the j^{th} row of **X**; column vector transpose will be denoted $(\mathbf{x}_i)^T$, italic characters will be used for scalars, e. g. matrix elements x_{ij} or indices *i*. From a given matrix **X** of *N* rows and *P* columns, the mean row, denoted by $\bar{\mathbf{x}}$, is calculated by $\bar{\mathbf{x}}^T = (\mathbf{1}_N^T \mathbf{1}_N)^{-1} \mathbf{1}_N^T \mathbf{X}_C$ where $\mathbf{1}_N$ is a column vector of *N* ones and the centered matrix will be denoted $\mathbf{X}_C$ and given by $\mathbf{X}_C = \mathbf{X} - \mathbf{1}_N \bar{\mathbf{x}}^T$.

3- Material and Methods

3.1- Experimental setup

The spectrometer was a “Bruker IFS 66v/S FTFIR”, equipped with a 23µm Mylar

beamsplitter which efficacy range covered from 150 to 20 cm^{-1} (0.6 – 4.5 THz). The source was a Hg lamp emitting a continuous radiation between 600 and 5 cm^{-1} (0.15 – 18 THz). The detector was a Si bolometer from “Infrared Laboratories Model” N° HD-3 which was cooled with liquid He at 4.2 K. It was equipped with a high-pass filter which covered the range between 0.5 to 10 THz. The diaphragm had an aperture of 12.0 mm wide. The resolution used has been set to 0.5 cm^{-1}. All measurements were under vacuum and at ambient temperature, 296 +/- 1 K. Reference has been registered every hour during the experiment, with a pure PE pellet. All spectra have been taken with 100 scans.

Samples were made up of ternary mixtures of PE powder, sucrose and glucose and of binary mixtures of PE and sucrose; and PE and glucose. Sucrose was obtained from Fluka Analytical (Ref. N° 84100) and glucose was purchased from Sigma-Aldrich (Ref. N° 16325). No further purification was needed for both products. Glucose powder of size <100 µm was used with no other preparation whereas sucrose has been grounded with a mortar and a pestle to reduce particle size (<500 µm). PE powder of size 100 µm was purchased from Aldrich (Ref. N° 26935-2, spectrophotometric grade powder). Sample discs were prepared by mixing different ratios of sugars with PE. Disc masses were of 70 ± 0.1 mg.

For ternary mixtures, the quantity of PE in discs was fixed at 70 % (m/m) of the total mass. The remaining 30 % (m/m) were composed of a mixture of glucose and sucrose, for a total of 21 samples. The first sample contained 0 % of glucose and 100 % of sucrose, the second 5 % of glucose and 95 % of sucrose and so on, with 5 % of mass changes, up to 100 % sucrose.

Binary mixtures were composed of PE powder and a single sugar. The masses used of glucose were: 1, 5*, 7, 9, 10, 13, 15*, 17 and 20* mg. The masses used for sucrose were: 1, 3*, 4, 5*, 7, 9*, 11, 13, 15*, 17, 18 and 20* mg. Repetition has been made for the

samples marked with asterisks.

All samples have been compacted in 0.5 ± 0.2 mm thick and 12 mm wide discs with a mean volume of 56.55 mm^3. Each disc has been grounded and compacted a second time to ensure homogeneity.

3.2- Data processing

Mean spectra were collected in transmission with the *Opus/IR* software. The spectra have been gathered in a matrix **T** of N rows (individual: sample spectrum) by P columns (variable: wavenumber). Matrix **T** has then been converted to the absorbance matrix **A** with the equation $\mathbf{A} = log_{10}(1/\mathbf{T})$ (Eq. 1). Samples with sucrose only and glucose only have been chosen as the calibration set. The mixture samples have been selected to be the test set. Outliers in the calibration were spotted and excluded by running Principal Component Analysis (PCA) with The Unscrambler software (CAMO, version 9.2). Samples which presented very high residual variance values or which were excluded by the T^2 limit of the Hotelling test have been removed from the calibration set.

The calibration models were created and tested with Matlab (version 7.4.0 R2007a). Partial Least Squares (PLS) regression with cross-validation, regression based on variable selection according to the CovSel technique [12], and regression based on Net Analyte Signal (NAS) have been tested [13].

Net spectra (NAS) of the two products glucose and sucrose have been extracted from absorbance spectra thanks to Eq. 10:

$\mathbf{K} = (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{A}$ (Eq. 10)

With **Y** the matrix regrouping the sugars concentration. According to the work of Faber [14, 15], multivariate sensitivity of the spectrometer to glucose and sucrose has been evaluated by the norm of NAS, i.e. **K** rows.

The standard errors mentioned in this paper (SECV, SEP_C) and bias have been calculated as follows:

SECV = RMSE ($\hat{y} - y$) (Eq.13);

SEP_C = RMSE ($\hat{y} - y - bias$) (Eq. 14);

Bias = mean ($\hat{y} - y$) (Eq. 15).

With RMSE the root mean square error, y the actual value of a sugar in a sample and $\hat{y}$ the predicted value of the sugar in the same sample.

4- Results and Discussion

The experimental spectra are shown in Fig. 1.

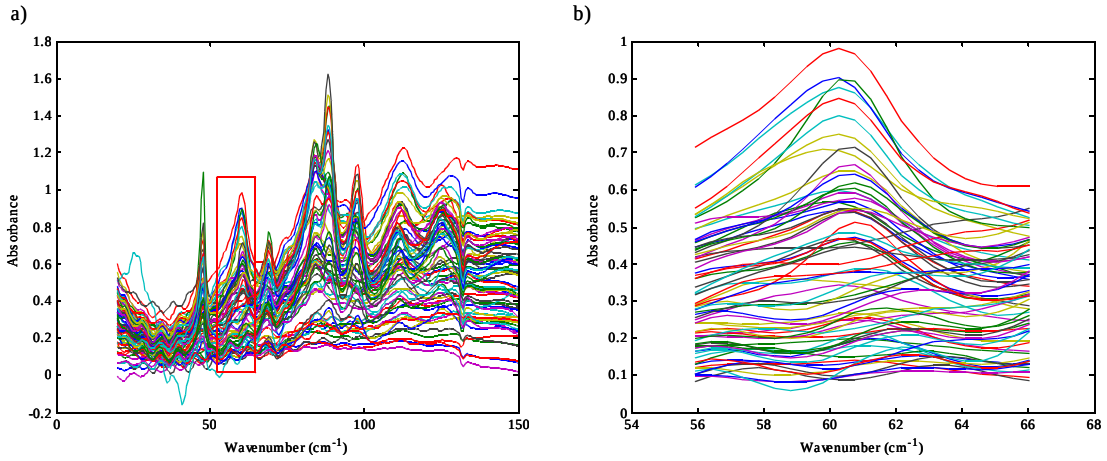

Fig. 1: Experimental spectra. Sucrose, glucose and mixtures have been acquired in absorbance. a) Spectra on the whole experimental range; b) Zoom on the 61 cm^{-1} peak. Zoom shows that spectra are not exactly organized according to the concentration in sugars.

The main characteristic features of glucose and sucrose can be seen on the spectra. Sucrose and glucose characteristic bands have already been identified, they can be found in online spectral database, such as Riken's [THz database Web: http://www.riken.jp/THzdatabase/ (Tera-photonics Laboratory, RIKEN Sendai)]. Sucrose peaks appear at 47.7, 60.2, 84.4, 95.4 and 111.4 cm^{-1}. Glucose peaks appear at 47.7, 69.4, 88.5, 99 and 125.8 cm^{-1}. A large feature next to 97 cm^{-1} appears only for mixtures; it

is believed that this peak is the combination of the peaks at 95.4 and 99 cm^{-1} because of their proximity. The feature appearing at 132 cm^{-1} is an artefact due to the beamsplitter and is not relevant of any information. Spectra present a strong baseline. This artifact is probably due to photon scattering in discs. This might be the consequence of the difference in the granulometry in samples which was not completely controlled and is known as the principal source of scattering.

4.1- Principal Component Analysis

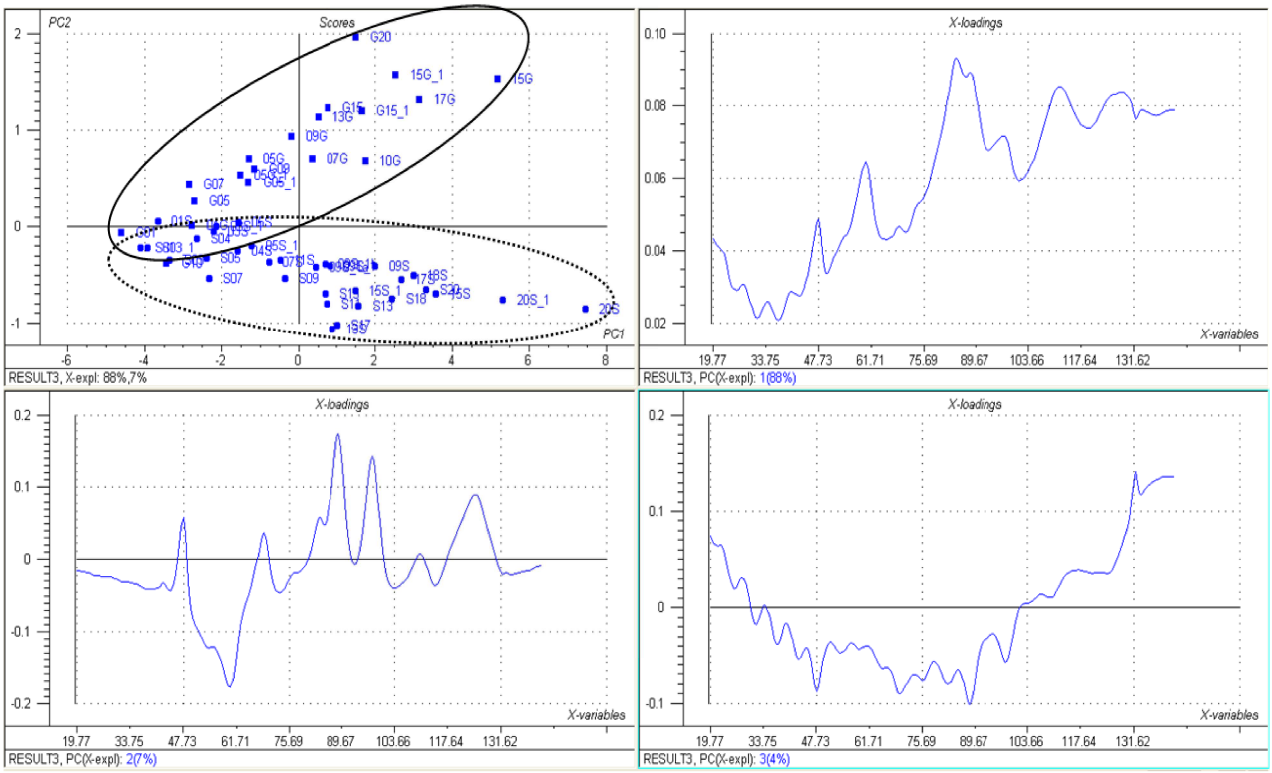

Fig. 2: PCA on calibration samples results. Upper left: scores on PC 1 and PC 2, sucrose samples are inside the circle in dotted lines, glucose samples are inside the circle in full line; upper right: loading on the 1st PC; down left: loading on the 2nd PC; down right: loading on the 3rd PC.

The PCA revealed the presence of a few outliers in the calibration set. They presented high variance residuals compared to the other samples and two of them added too much leverage to the set. Figure 2 shows the results of the PCA after the removal of these outliers. Scores for two first components and the loadings for the first three principal components (PC) have been displayed.

Scores reveal the nature of the two first LVs. First PC is related to the concentration of sugar inside the pellets. Moreover, the relation between the scores and sugars showed that sucrose is mainly related to the first PC, while both first and second PCs contribute to the distribution of glucose. The first loading explains 88 % of the total variance. It presents all the characteristic features of glucose and sucrose: glucose at 47.7, 69.4, 88.5 and 125.8 cm^{-1}, sucrose at 47.7, 60.2, 84.4 and 111.4 cm^{-1}. It is believed that the peaks present at 99 cm^{-1} for glucose and at 95.4 cm^{-1} for sucrose appear as combined in a single feature at 97.4 cm^{-1}. Thus, the first loading concerns the variation detected between sugars. The second loading only explains 7 % of the total variance, but its features are present at the same locations where the main bands of sucrose and glucose vary the most. The third loading only explains 4 % of the explained variance but it presents two notable features. Firstly, the features appearing at 47.7, 69.4, 88.5 and 97.8 cm^{-1} correspond to glucose features only. We deduced that sucrose was mainly explained by the two first PCs. On the contrary, the analysis of glucose requires more components to be done. Secondly, the sinusoid that has been spotted on the spectra is found again in the third loading. This structured noise can be identified easily with PCA, which could be a starting point if one would want to take care of it. We chose not to remove the noise, as it does not interfere with the zones which contain data information. So, the calibration methods that have been selected should not be influenced (or will not take in consideration) by this sinusoid. However, solutions exist if we would have wanted to remove noise: spot the frequency corresponding to the noise and remove it from the interferogram, realize an orthogonal projection to the noise matrix in order to cancel its effects, etc. [16]

4.2- Partial Least Squares Regression

4.2.1- Calibration

According to our previous work [17], we have shown that the spectral extremities are not areas of interest. With our configuration, we have recommended discarding the whole spectral range 20 – 150 cm^{-1} as the extremities showed little repeatability and artefacts that are not due to the sample. We suggested using only the variables between 50 and 130 cm^{-1}. To confirm the utility of this variable selection, we compared firstly the PLS models for sucrose and glucose obtained on the whole spectral range and on the reduced range. The results are shown in Table 1.

	Sugar	LVs	R^2	Bias (g/L)	SECV ($g.L^{-1}$)
20-150 cm^{-1}	Sucrose	6	0.911	2.27	36.6
	Glucose	6	0.931	-0.34	26.8
50-130 cm^{-1}	Sucrose	8	0.937	-1.29	30.3
	Glucose	7	0.939	-0.792	24.9

Table 1: PLS 1 models for glucose and sucrose prediction. Comparison between entire spectral range and reduced range.

The models created with the entire spectral range need fewer principal components to work. Only 6 Latent Variables (LVs) are used to propose a model for glucose and for sucrose. However, the models created with the reduced spectral range only need one more LV for glucose and two more LVs for sucrose to propose better models. As there is only two chemicals (glucose and sucrose), a model with two LVs could be sufficient to interpret the model, but including six LVs or more is needed to remove most of the non-chemical effects of the models (baseline, etc.). The coefficients of determination slightly increase for both sugars: from 0.911 to 0.937 for sucrose and from 0.931 to 0.939 for glucose. For both sugars nevertheless, the Standard Error of Cross Validation for the calibration set (SECV) has been lowered when using the reduced spectral range.

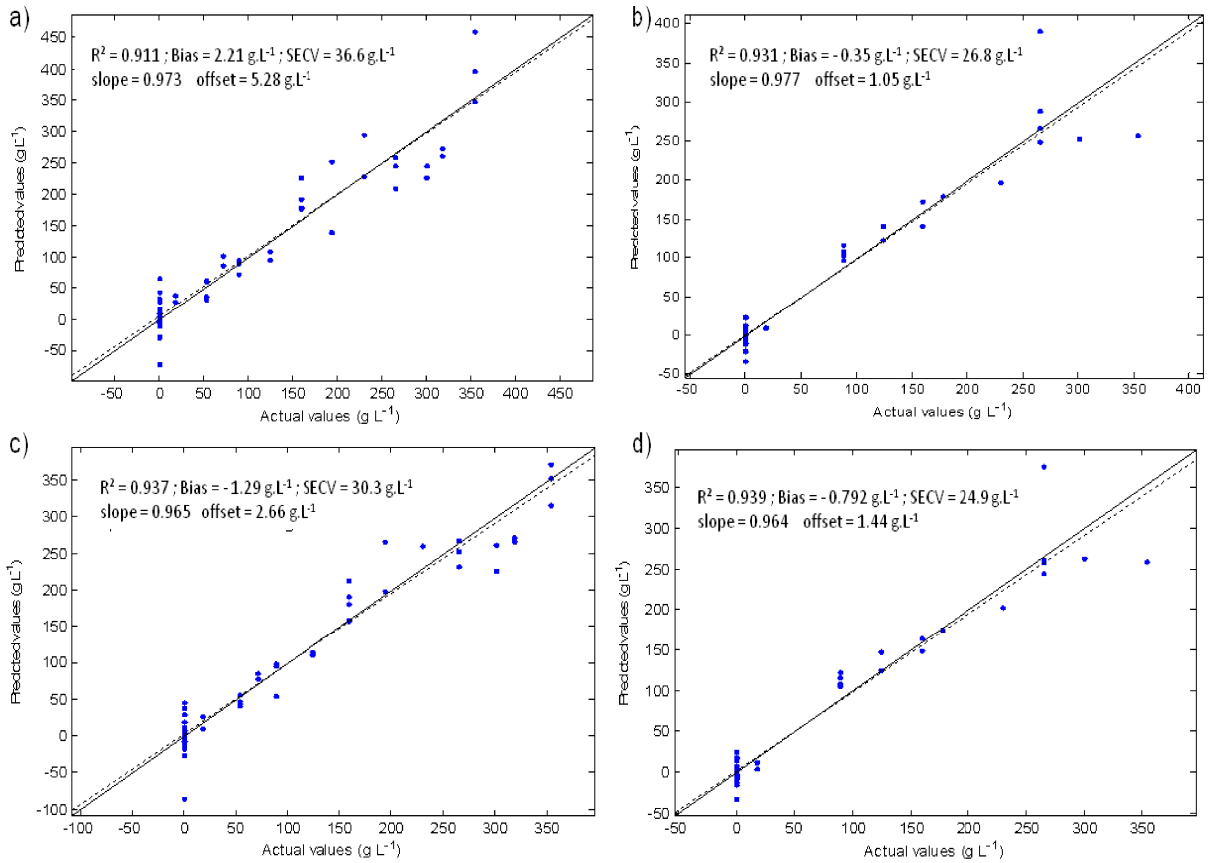

Fig. 3: PLS 1 models with cross validation on raw data. a) Sucrose model for entire spectral range; b) glucose model for entire spectral range; c) sucrose model for reduced spectral range; d) glucose model for reduced spectral range.

Figure 3 presents the created PLS models obtained with cross validation and without any pretreatment. As the previous values have shown it, the models with the entire range and with reduced range are quite similar. However, the repetition samples are closer in the model with reduced spectral range: there is less variability between their predictions, which improves the models. For these reasons, the spectra have been reduced to the 50 – 130 cm^{-1} range before applying pretreatments.

On this figure, one can also see that there is a great variation in the prediction of the samples actually containing 0 $g.L^{-1}$ of one sugar (i.e. all glucose samples in the sucrose model and vice versa). The proximity of the characteristic features of glucose and sucrose causes the models to misinterpret some data of one sugar for another while there is no common features at the exception of the peak at 48 cm^{-1}. Even after the wavenumber selection, which excludes this peak, the dispersion subsists, so the hypothesis of the

dispersion caused by a sole feature can be discarded. This misinterpretation is linked to the whole spectra, so a factor of selection for the calibration models has to be the reduction of the dispersion of the samples containing none of the sugar concerned by the model.

In order to optimize the calibration, it is needed to reduce the importance of additive and multiplicative effects that could be present in our samples spectra. A second derivative of Savitzky-Golay has been applied to raw spectra in order to treat additive effects [18-20]. The PLS models that have been computed are shown in Figure 4.

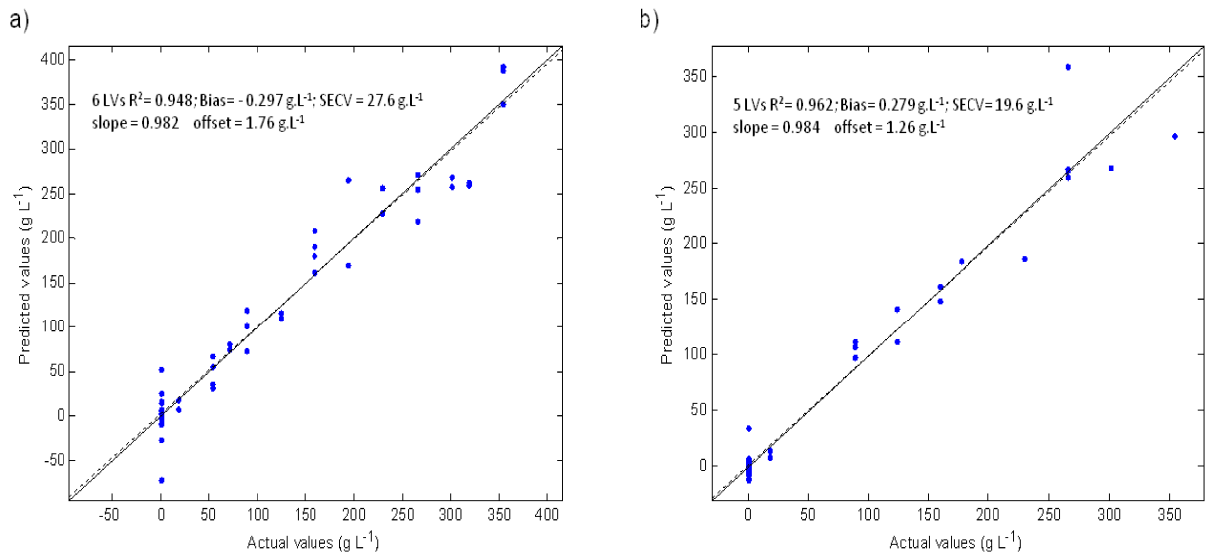

Fig. 4: PLS calibration models with cross validation, with second derivative for a) sucrose; b) glucose.

The first observation is that the number of LVs used for these models has diminished when compared to the models obtained with raw data. It has been lowered from 8 to 6 LVs for sucrose and from 7 to 5 LVs for glucose. R^2 has been lightly improved. It is also the case for SECV which varies from 30.3 $g.L^{-1}$ to 27.6 $g.L^{-1}$ and from 24.9 $g.L^{-1}$ to 19.6 $g.L^{-1}$ for sucrose and glucose respectively. This effect can be seen on Figure 4, where the predicted values of the samples are closer to the actual values. Actually, the points with a volumetric concentration of 0 $g.L^{-1}$ of sugar are less distant one from another. The use of the second derivative from Savitzky-Golay on raw data allowed the improvement of the prediction models because it reduced the baseline effect that can be seen on Figure 1. It

shall be of interest to note that we could have used a model for sucrose with 2 LVs only: this model would have been under-fitted but it would have presented a low dispersion of glucose samples predicted at 0 $g.L^{-1}$. We could also have tested a model with 7 LVs for sucrose, but this model would have been over-fitted, a better model in itself, but with a wide dispersion of glucose samples. Considering the variance of the model, we finally opted for a model with 6 LVs that we discussed previously in this article.

As a complement for second derivative, a Standard Normal Variate (SNV) pretreatment has been applied in order to remove multiplicative effects. The created models are shown in Figure 5.

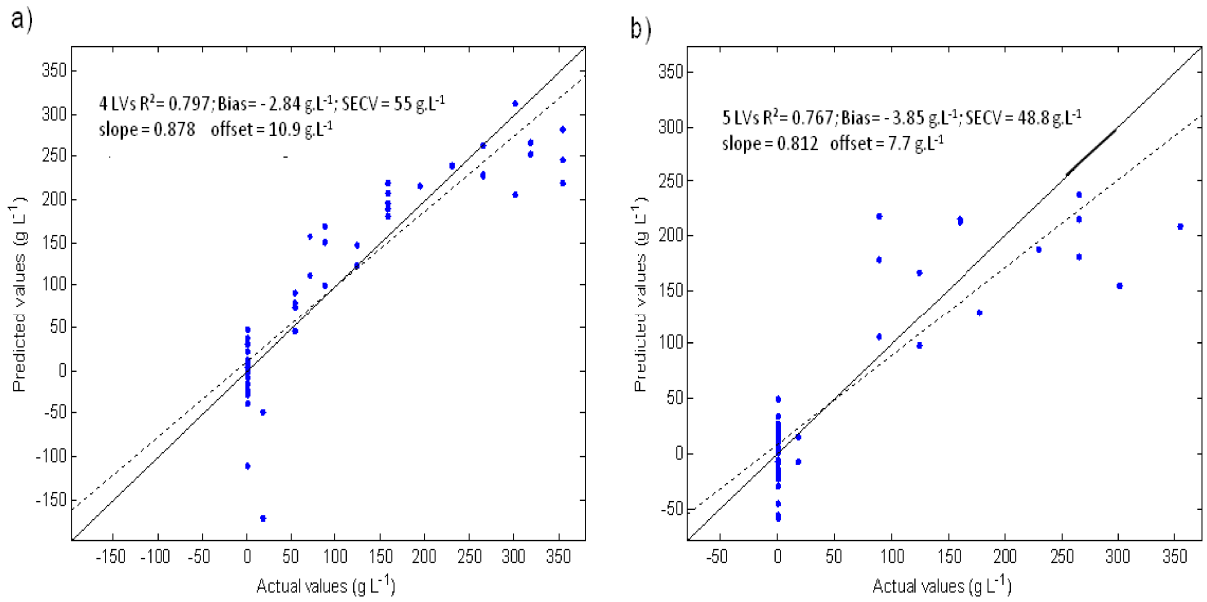

Fig. 5: PLS models with cross validation, with second derivative and SNV for a) sucrose; b) glucose.

The addition of SNV greatly disturbed the models. The determination coefficients of the two models are greatly diminished, from 0.95 to 0.78. The number of conserved LVs has diminished for both sugars, the utilisation of second derivative and of SNV resulting in the diminution of the number of components required to explain the model. However, the biases of the models are more important. It is the same for SECV that are superior to those determined for the models based on raw spectra. SNV has to be used only if a multiplicative effect is not directly connected to data information. In the case of our study,

the absorbance of spectra is related to the sugar concentration of the pellets. Thus, the global level of the spectrum contains the information; there is no multiplicative effect to remove. The use of SNV has dilated spectra: during normalization, it improved the importance of the zones with few or no information while it reduced the importance of the zones possessing the information, making the models unusable for quantification.

4.2.2- Models testing

The best PLS1 model developed being the one pretreated with only a second derivative, test set (mixtures spectra) is applied for both sugars. The test results are shown Figure 6.

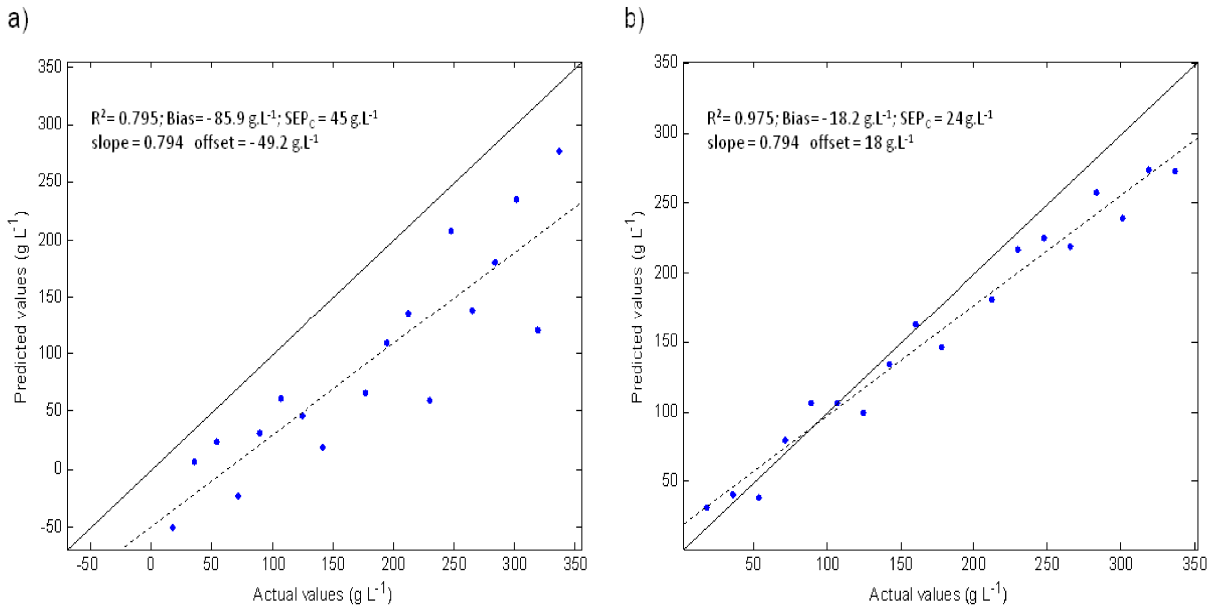

Fig. 6: Test of PLS models on mixtures. a) Quantity of sucrose in mixture pellet; b) quantity of glucose in mixture pellet.

The test shows that the models created with PLS regression possess strong biases and slopes. While the use of second derivative has improved the calibration models, the prediction of sucrose and glucose concentrations in mixtures is not precise. The predictions are much less precise for sucrose than for glucose. The determination coefficient of glucose is closer to 1 than the coefficient for sucrose, the bias and the Standard Error of Prediction (SEP_C) for the test are less important. This is confirmed by Figure 6 where the sucrose concentrations predicted in mixtures are strongly dispatched

far from the regression line (hence SEP_C = 45 $g.L^{-1}$), while the glucose concentrations are better divided around their regression line, which explains the SEP_C of 24 $g.L^{-1}$. Other pretreatments have to be tested in order to determine if they can improve the test set prediction.

4.3- CovSel variable selection

The CovSel technique is a new developed variable selection technique. It distinguishes from the others variable selection techniques because it takes in account that all spectra variables are correlated, which means that it realises a step by step selection of variables regarding their covariance with all the responses, and that it performs an orthogonal projection to the selected variable to integrate complementary and structured data [12]. This technique is also able to handle many responses. The results presented with the CovSel method concern only raw spectra as we focused on the feasibility of its use. Spectra pretreated with second derivative before using CovSel have been tested, but the model was not improved and thus not judged of interest in this paper. Figure 7 presents the selected variables and the calibration models that have been created.

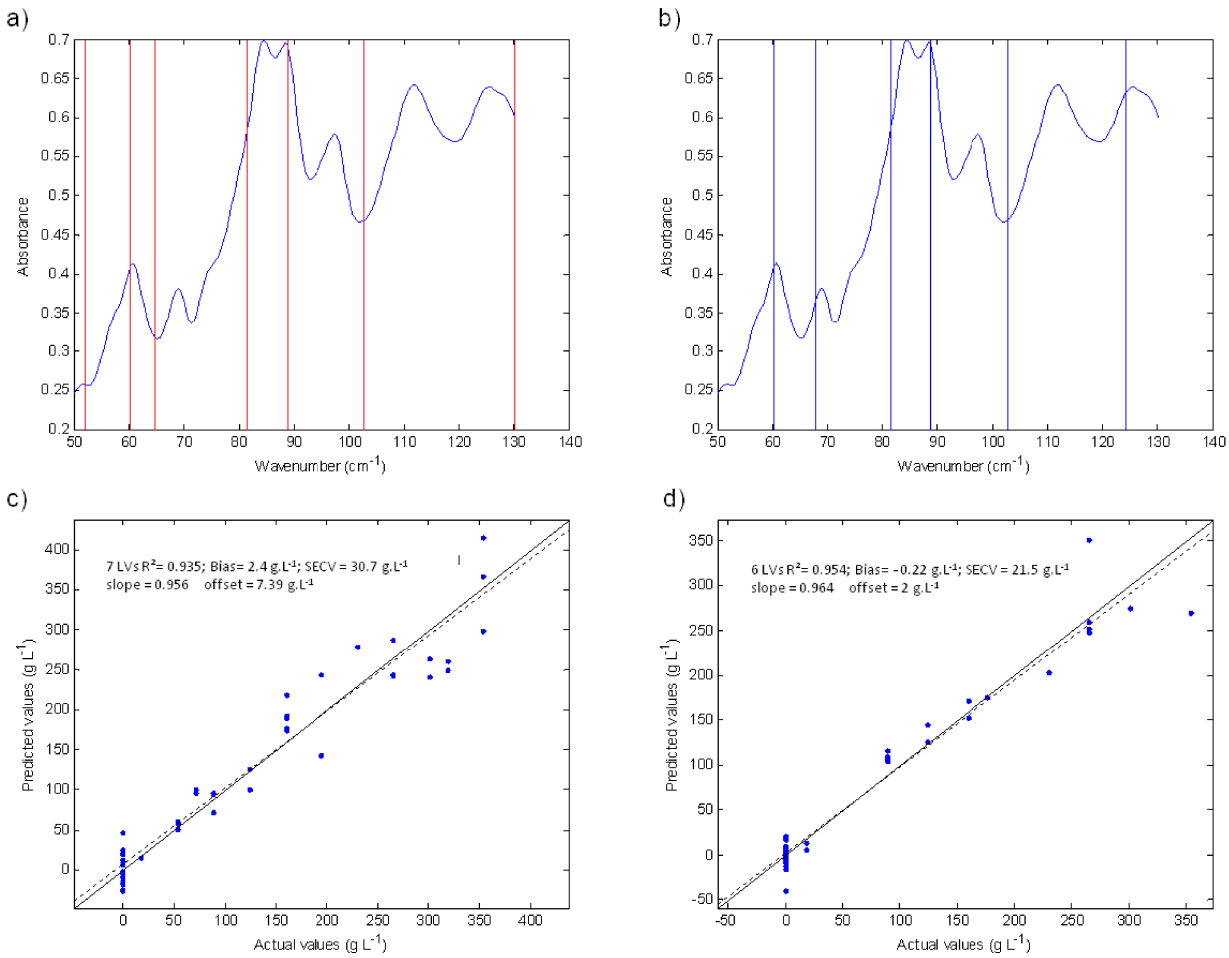

Fig. 7: CovSel-based prediction models. a) Wavenumbers selected for sucrose on the mean spectrum of all spectra of the calibration set; b) wavenumbers selected for glucose; c) prediction model for sucrose; d) prediction model for glucose.

Figure 7 a) presents the seven wavenumbers selected for sucrose analysis. The wavenumbers at 52.1, 64.6, 102.7 and 130 cm^{-1} do not correspond to sucrose characteristic features but are located in areas with strong variance. The wavenumber at 60.2 cm^{-1} correspond to one of the sucrose peaks, and the wavenumber selected at 81.5 cm^{-1} is close to the band present at 84.4 cm^{-1}. It shall be noted that the last selected wavenumber is at 88.5 cm^{-1}, which correspond to the position of a glucose peak. The five selected variables for glucose are shown Figure 7 b). Identically, the wavenumber at 102.7 cm^{-1} does not correspond to a characteristic peak but to a zone located just before a common band to glucose and sucrose with strong variation between spectra. Wavenumbers at 60.2 and 81.5 cm^{-1} have already been mentioned for sucrose; the variables at 68, 88.7 and 124.4 cm^{-1} correspond to glucose bands. The models created by

this variable selection are pretty much equivalent, when comparing their characteristics, to the models created when performing a second derivative of Savitzky-Golay on raw spectra, although the SECV in the last case are slightly better (sucrose: 30.7 $g.L^{-1}$ for 27.6 $g.L^{-1}$, glucose: 21.5 $g.L^{-1}$ for 19.6 $g.L^{-1}$). Figures 7 c) and d) reveal that the dispersion of the samples containing 0 $g.L^{-1}$ of one sugar is a little more important than in previous models. However, the test of the models obtained by CovSel variable selection is also based on the covariance of the selected variables with the responses, which results in the predictions shown in Figure 8.

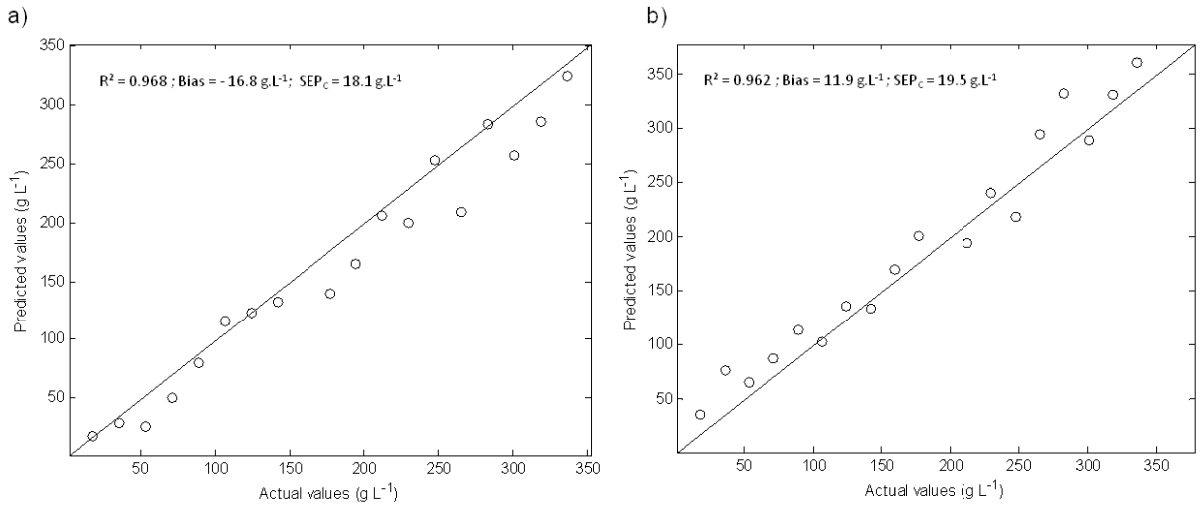

Fig. 8: Test for CovSel models. a) sucrose; b) glucose.

Figure 8 a) shows that the prediction of the concentration in sucrose in the mixture pellets has been greatly improved with this model than with the model where the second derivative was applied. The SEP_C has diminished from 45 to 18.1 $g.L^{-1}$. Bias has also improved from -85.9 to -16.8 $g.L^{-1}$. The determination coefficient is really close to one, which was not the case. An improvement is also noticed for glucose prediction as it is shown in Figure 8 b). The SEP_C for the CovSel model is equal to 19.5 $g.L^{-1}$ while it is equal to 24 $g.L^{-1}$ for the PLS model with a second derivative, while bias and R^2 are equivalent. Actually, even if the models created by variable selection looked like less performing than the models created by PLS at first glance, the results of the quantification of sugars in

mixtures are clearly better.

4.4- Net Analyte Signal (NAS)

NAS allow recovering net, ideal spectra, from an experimental spectral matrix. Their determination has been the object of other publications [13-15, 17]. As a matter of fact, it is possible to obtain the most optimal spectra in terms of unmixing. So, the prediction on sucrose and glucose samples (Figure 9), then the prediction on mixtures (Figure 10) can be achieved. As we have demonstrated that the spectra pretreated with a second derivative of Savitzky-Golay allowed better models, we chose to create the NAS with pretreated spectra rather than raw spectra.

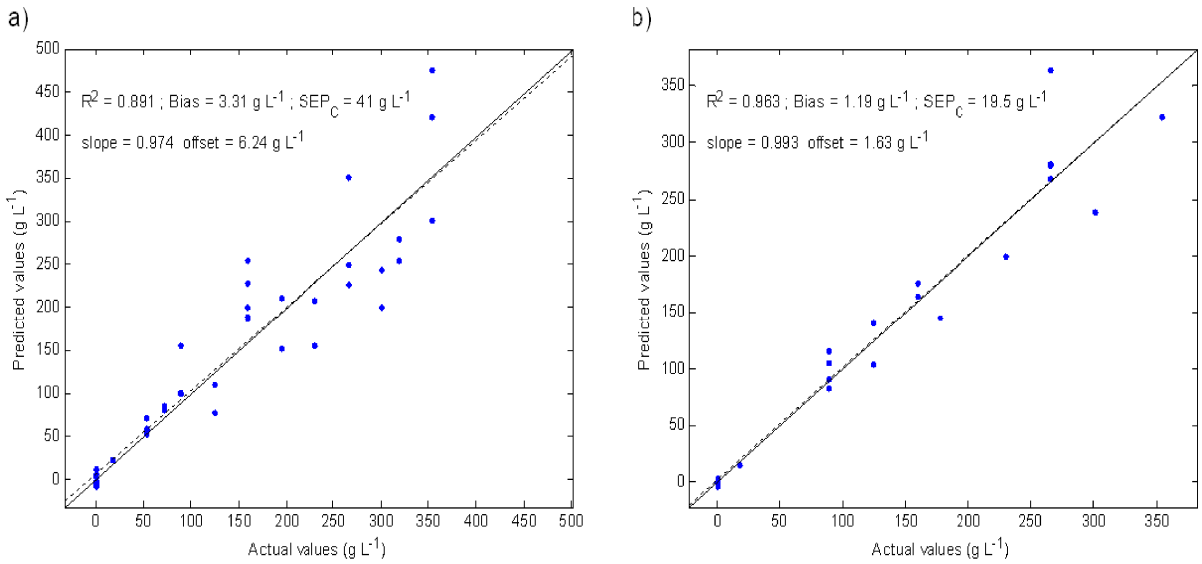

Fig. 9: Projection of calibration samples on NAS. a) Prediction model for sucrose; b) Model for glucose.

Figure 9 a) shows that the sucrose samples have got a very low bias when projected on NAS. However, their dispersion is still important when compared to other methods: SEP_C equals 41 $g.L^{-1}$. On the other hand, the glucose prediction model shown in Figure 9 b) is the best model we have created. The determination coefficient is 0.963, the bias is low and SEP_C equals 19.5 $g.L^{-1}$. The prediction when projecting the spectra of sucrose and glucose only on NAS is better than the predictions calculated for the other pretreatment methods.

Moreover, one shall note that when using this method, the dispersion of the samples containing 0 $g.L^{-1}$ of a sugar is minimal: the prediction of those samples is close to 0. There is no more an effect which could impact on predicted concentrations when using the NAS. This is due to the explicit orthogonalization that happens for a product towards the other when using this method: the products become totally independent; there is not interaction between them.

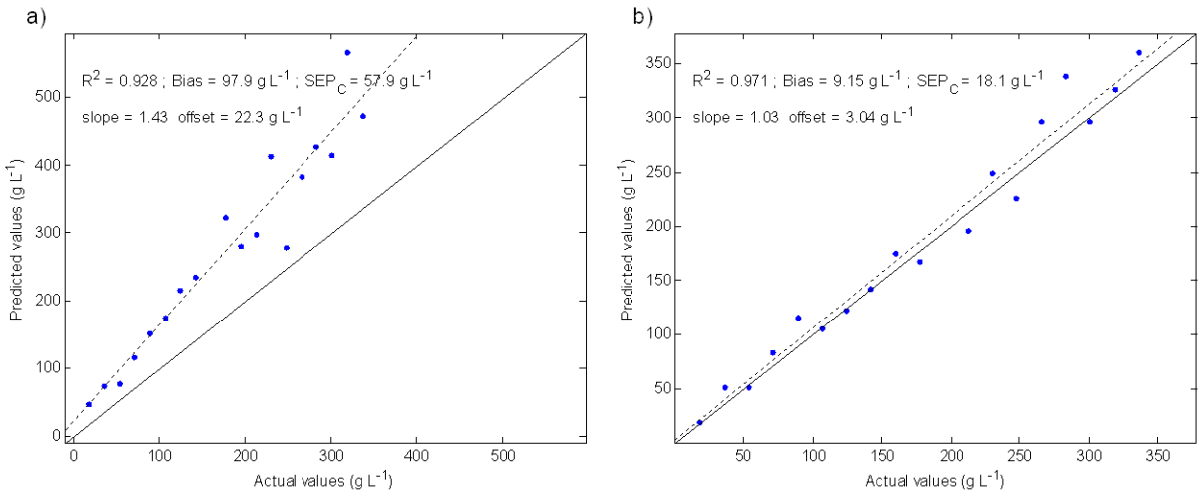

Fig. 10 : Projection of mixtures on NAS. Prediction model for a) sucrose; b) glucose.

The result of the quantification of the mixtures with NAS is less convincing. In both cases, strong biases exist, especially for sucrose prediction (97.9 $g.L^{-1}$). These biases can nevertheless be treated by a bias-slope correction to improve the quality of the models. The error of prediction for the sucrose concentration in mixtures is higher for strong concentrations than for the lower ones. It is characterised by the stronger value of SEP_C: 57.9 $g.L^{-1}$. On the contrary, Figure 10 b) reveals that the glucose concentration quantification is quite precise, for high concentrations as well as low concentrations, with a SEP_C of 18.1 $g.L^{-1}$. The bias for glucose can also be treated with a bias-slope correction.

5- Conclusion

We have shown that chemometrics can be transposed without major drawback to THz spectroscopy for sugar contents assessments. We have determined that a PLS model created after the derivation of spectra is not the better prediction model for the quantification of sugars in mixtures; even while SECV is low. Moreover, the adjunction of SNV (which is sometimes used automatically when pretreating spectra to deal with potential additive effects) has led to an undesired effect. The model crashed because of the normalization induced by SNV. When the spectral information itself is the source of a multiplicative effect on spectra, pretreating those spectra with SNV removes this information. Therefore, SNV is useless in such case. Other chemometric methods furnish equivalent calibration models but better predictions. The CovSel method is suited for both sugars and gives the best overall results: SECV for sucrose (30.7 $g.L^{-1}$) and for glucose (21.5 $g.L^{-1}$) are a little higher than for other models, but they are more efficient for the test on mixture samples. Finally, the NAS utilisation for quantification may be an interesting lead to follow. While this technique is working nicely on glucose samples, we have shown that it can be improved for sucrose quantification with a bias-slope correction. However, the prediction error in the sucrose model for samples with a high concentration can be a crippling factor. This technique is nevertheless the only one which allowed a significant decrease of the dispersion of the samples containing 0 $g.L^{-1}$ of a sugar (for both models), so the discrimination of these products is more efficient. Therefore, NAS could be used to ease the distinction and the quantification of different products which spectra are more or less similar. Other studies will be realised on more complex mixtures to enrich the catalogue of chemometrics methods used for calibration in THz spectroscopy and to confirm that this analytical technique conserve a good capacity for quantifying the different constituents of mixtures.

REFERENCES

1. B. Fischer, M. Hoffmann, H. Helm, G. Modjesch, P.U. Jepsen, Semicond. Sci. Technol. 20 (2005) S246–S253.

2. J.F. Federici, B. Schulkin, F. Huan, D. Gary, R. Barat, F. Oliveira, D. Zimdars, Semicond. Sci. Technol. 20 (2005) S266-S280.

3. J. Hooper, E. Mitchell, C. Konek, J. Wilkinson, Chem. Phys. Lett. 467 (2009) 309-312.

4. D.G. Allis, J.A. Zeitler, P.F. Taday, T.M. Korter, Chem. Phys. Lett. 463 (2008) 84-89.

5. G.F. Liu, X.J. Ma, S.H. Ma, H.W. Zhao, M.W. Ma, M. Ge, W.F. Wang, Chinese Journal of Chemistry 26 (2008) 1257-1261.

6. Y. Li Hor, H.C. Lim, J.F Federici, E. Moore, J.W. Bozzelli, Chemical Physics 353 (2008) 185-188.

7. P.F. Taday, R. Soc. Lond. A 362 (2004) 351-364.

8. M. Ge, H. Zhao, T. Ji, X. Yu, W. Wang, W. Li, W. Science in China: Series B Chemistry 49 (2006) 204-208.

9. M. Walther, B.M. Fischer, P.U. Jepsen, PU. Chem. Phys. 288 (2003) 261-268.

10. H. Wu, E.J. Heilweil, A.S. Hussain, M.A. Khan, Journal of Pharmaceutical Sciences 97 (2008) 970-984.

11. C.J. Strachan, P.F. Taday, D.A. Newham, K.C. Gordon, J.A. Zeitler, M. Pepper, T. Rades, Journal of Pharmaceutical Sciences 94 (2005) 837-846.

12. J.M. Roger, B. Palagos, D. Bertrand, E. Fernandez-Ahumada, Chemometrics and Intelligent Laboratory Systems 106 (2011) 216-223.

13. H. Martens, T. Naes, *Multivariate Calibration* (John Wiley & Sons Ltd, Chichester, 1989).

14. A. Lorber, K. Faber and B. R. Kowalski, Anal. Chem. 69 (1997) 1620-1626.

15. N. K. Faber, Anal. Chem. 71 (1999) 557-565.

16. J. M. Roger, F. Chauchard, V. Bellon-Maurel, Chemometrics and intelligent laboratory systems 66 (2003) 191-204.

17. M. Papillaud, J. M. Roger, F. Davrieux, submitted / unpublished work.

18. A. Savitzky, J. E. Golay, Anal. Chem. 36 (1964) 1627-1639.

19. H. H. Madden, Anal. Chem. 50 (1978) 1383-1386.

20. P. A. Gorry, Anal. Chem. 62 (1990) 570-573.

Conclusion générale

Le but du travail présenté dans cette thèse a été de déterminer la faisabilité de la quantification d'un mélange de poudres par spectroscopie Terahertz au sein d'une matrice transparente ; avec l'objectif d'évaluer le potentiel de cette technique pour le contrôle non destructif des produits agroalimentaires. Les fruits et légumes sont des systèmes biologiques extrêmement complexes. La technique et le matériel actuels ne permettent pas de caractériser un tel milieu tel quel. L'eau absorbant très fortement les rayonnements THz, il serait nécessaire de la retirer du système que nous souhaitions étudier. Un fruit déshydraté serait constitué principalement de cellulose (transparente au THz), de sucres, d'amidon, de vitamines, etc. C'est un mélange encore trop complexe pour l'étude ; nous avons donc dû procéder à la simplification du système en remplaçant un fruit par une matrice équivalente. Le mélange de constituants a également été simplifié pour étudier un mélange de deux sucres, le glucose et le sucrose, en prélude à l'étude de systèmes plus complexes.

L'analyse de la problématique a conduit à dégager trois grandes étapes dans notre recherche : l'état de l'art sur la spectroscopie Terahertz dans les domaines chimiques et biologiques ; l'étude métrologique de l'équipement utilisé ; la faisabilité de la quantification de produits simples dans une matrice transparente. Pour chacune de ces étapes, un article a été soumis pour publication.

Nous avons vu que la spectroscopie THz est un domaine en pleine évolution. Les publications actuelles sont principalement basées sur l'analyse de produits et de systèmes simples. Des études portant sur des systèmes plus complexes commencent à faire leur apparition. Par ailleurs, l'industrie pharmaceutique est actuellement l'un des principaux utilisateurs de la spectroscopie THz en tant que complément, voire en remplacement de certaines techniques spectroscopiques actuelles. L'un des intérêts majeurs de cette méthode est qu'elle agit au niveau des vibrations intermoléculaires et va donc fournir une véritable empreinte digitale des molécules. Dans le cadre de la caractérisation des produits, cette étape peut se révéler très importante pour la distinction entre deux produits proches

structurellement. D'autre part, les rayonnements THz étant non ionisants et non invasifs, ils présentent l'avantage de ne pas être dangereux pour l'intégrité du système étudié.

Lors de l'analyse métrologique de l'appareil, celle-ci a permis de mettre en évidence un défaut d'acquisition du spectromètre. Celui-ci se traduisait par un bruit sinusoïdal additionné aux spectres. L'identification et la compréhension de ce phénomène ont été facilitées par l'utilisation de techniques multivariées. Nous avons également montré que la sensibilité de l'appareil n'était, a priori, pas la même pour les deux sucres étudiés. Nous pouvons émettre l'hypothèse que cela dépend fortement de l'environnement immédiat des molécules (d'un point de vue chimique) et de l'importance des vibrations intermoléculaires qui sont en jeu. Il en découle qu'au sein d'un milieu complexe, certaines molécules (présentant peu de groupes capables d'établir ce genre de liaisons) ne seront pas aussi facilement détectées que d'autres. Il devient nécessaire d'augmenter d'une manière ou d'une autre la sensibilité de l'appareil à ces molécules afin de pouvoir réaliser une étude complète dans un milieu complexe. Nous avons pu montrer que certaines techniques chimiométriques, comme le Net Analyte Signal, permettent de mettre les signaux de nos deux produits au même niveau. La détection et la distinction des produits deviennent plus aisée.

Les premières études sur un mélange de poudres ont montré que la spectroscopie THz se prête naturellement à l'utilisation de méthodes chimiométriques en vue d'applications de quantification. La connaissance que nous avons produite à ce sujet est d'ordre quantitatif. Nous avons en effet pu montrer qu'il est possible de prédire les concentrations en sucres au sein d'un mélange à partir d'un étalonnage réalisé sur des échantillons purs uniquement. Les différents prétraitements utilisés et les modèles créés sont de qualité variable, et nous avons montré que l'utilisation de modèles chimiométriques de base tels que la PLS (avec des spectres prétraités ou non) ne suffisent pas à prédire de manière satisfaisante la concentration en sucres dans une pastille. Par contre, l'utilisation de nouvelles méthodes, comme la sélection de variables CovSel ou l'application à nouveau des NAS, permet d'affiner ces modèles et d'améliorer la quantification. Jusqu'ici, seule la quantification

de produits uniques ou la proportion entre deux formes structurelles d'une même molécule avait été réalisée ; notre travail présente une nouvelle approche contribuant à l'analyse de milieux plus complexes.

Nous avons montré le potentiel de la spectroscopie THz pour la caractérisation et la quantification des molécules dans des mélanges simples. La première étape de recherche dans la continuité de notre travail sera de complexifier au fur et à mesure les milieux sur lesquels nous travaillons en augmentant la quantité de molécules dans les mélanges. L'absorption de l'eau étant un problème, il sera nécessaire de l'introduire dans les systèmes dans l'optique de travailler directement sur les fruits et légumes, et non sur une matrice déshydratée. Des travaux ont ouvert la voie dans ce domaine et traitent de la caractérisation et de la quantification d'alcool dans des solutions aqueuses [Jepsen 2007]. Le matériel utilisé est également un facteur à prendre en compte. L'utilisation d'autres sources ou détecteurs orienterait la recherche de manière légèrement différente. Par exemple, nous avons démontré que la sélection de variables CovSel donne de meilleurs résultats en termes de quantification des sucres. Si nous souhaitions approfondir cette étude en nous focalisant sur les variables sélectionnées par l'algorithme, nous pourrions utiliser une source constituée d'un laser dont la fenêtre optique est beaucoup plus réduite que celle qui a été utilisée lors de la thèse, fournissant un rayonnement plus énergétique et pouvant potentiellement fournir plus d'informations, abaissant par la même occasion la limite de détection de la méthode. Nous pouvons également signaler que, dans un futur proche, la miniaturisation du matériel THz permettrait de franchir une nouvelle étape dans les applications de la spectroscopie THz et pourrait lui conférer la même diversité et la même aura que possèdent d'autres techniques spectroscopiques non destructives.

D'après nos observations, nous estimons que la mesure de résidus de pesticides par spectroscopie THz n'est pas encore opérationnelle. Le seuil de détection de la méthode (dans les conditions expérimentales utilisées) n'a pas été déterminé lors de nos expériences, aussi nous ne pouvons pas affirmer si des résidus de pesticides peuvent être mesurés ou si une concentration minimale

supérieure aux limites maximales de résidus (LMR) énoncées par l'Union Européenne est nécessaire. De plus, l'étude de quantification a montré que les modèles ne sont pas assez précis par rapport aux exigences de l'Union Européenne. La reproductibilité de la méthode n'étant pas non plus assurée, la spectroscopie THz n'est pas encore un outil adapté à la quantification de pesticides. Par contre, son mode de vibration en fait un outil de caractérisation idéal. Il est toutefois nécessaire de se constituer une base de connaissances de pesticides. Ceux-ci n'ont pas forcément un spectre bien défini (cf. Annexes) car les molécules chimiques (organo-chlorés, organo-phosporés…) qui les composent ne possèdent pas nécessairement les groupements chimiques qui sont responsables des vibrations inter-moléculaires caractéristiques de la plage spectrale THz. Néanmoins, il se pourrait qu'en étendant la gamme spectrale du THz vers l'IR, l'information contenue dans les spectres permettent d'identifier les différents pesticides, même s'ils n'ont pas un spectre caractéristique dans la seule bande THz. Comme nous l'avons signalé plus tôt, la présence d'eau constitue également un challenge pour les mesures THz. La question des mesures de résidus sur des fruits ou légumes (donc majoritairement composés d'eau) reste entière. Deux possibilités sont envisagées : premièrement la mesure par transmission de coupes d'échantillons de quelques microns d'épaisseur, auquel cas il sera nécessaire de trouver un moyen de gérer l'eau contenue dans les cellules, mais qui va à l'encontre de l'objectif initial de développer une méthode d'analyse non invasive ; et deuxièmement la mesure des échantillons par réflexion, où seuls les pesticides en surface seront analysés, mais la puissance du signal sera éventuellement diminuée par ce mode de mesure. Dans ce dernier cas de figure, l'eau ne risque plus d'interagir avec la mesure, mais se pose alors la question de la présentation de l'échantillon au rayonnement THz : les mesures ne peuvent plus être faites sous vide, elles devront avoir lieu depuis l'extérieur du spectromètre et il n'est pas possible de s'assurer du parallélisme des faces entre les fenêtres de mesures et l'échantillon. Au vu du développement des émetteurs et des détecteurs, on peut penser que les limitations actuelles seront dépassées d'ici quelques années et que le dosage de pesticides en laboratoire sera possible, de plus

amples études seront toutefois nécessaires pour envisager l'analyse de pesticides dans les champs de culture, surtout au point de vue miniaturisation du matériel dans le but de concevoir des spectromètres embarqués et dont les spectres pris à atmosphère ambiante et dans des conditions expérimentales variables (température, pression atmosphérique, humidité de l'air...) soient exploitables.

Références Bibliographiques

D.G. Allis, J.A. Zeitler, P.F. Taday, T.M. Korter, Chem. Phys. Lett. 463 (2008) 84.

D.H. Auston, Applied Physics Letters 26 (1975) 101.

D.H. Auston, P. LeFur, Appl. Phys. Lett. 28 (1976) 21.

D. Bertrand, E. Dufour, *La spectroscopie infrarouge et ses applications analytiques* Tec & Doc, Ed. (Lavoisier, Paris, 2000), 2nd ed., p. 7.

R. Bro, C.M. Andersen, J. Chemom. 17 (2003) 646.

R.P. Cogdill, R.N. Forcht, Y. Shen, P.F. Taday, J.R. Creekmore, C.A. Anderson, J.K. Drennen III, J. Pharm. Innov. 2 (2007) 29.

D.A. Crawley, C. Longbottom, B.E. Cole, C.M. Ciesla, D. Arnone, V.P. Wallace, M. Pepper, Caries Res. 37 (2003) 352.

A. Crocker, H.A. Gebbie, M.F. Kimmit, L.E.S. Mathias, Nature 201 (1964) 250.

A.G. Davies, A.D. Burnett, W. Fan, E.H. Linfield, J.E. Cunningham, Materials today 11 (2008) 18.

D. Dragoman, M. Dragoman, Progress in Quantum Electronics 28 (2004) 1.

N. K. Faber, Anal. Chem. 71 (1999) 557.

J. Faist, F. Capasso, D.L. Sivco, C. Sirtori, A.L. Hutchinson, A.Y. Cho, Science 264 (1994) 553.

J. Faist, F. Capasso, C. Sirtori, D.L. Sivco, J.N. Bailaargeon, A.L. Hutchinson, A.Y. Cho, Appl. Phys. Lett. 68 (1996) 3680.

J. Faist, F. Capasso, D.L. Sivco, A.L. Hutchinson, A.Y. Cho, Appl. Phys. Lett. 72 (1998) 680.

J.F. Federici, B. Schulkin, F. Huan, D. Gary, R. Barat, F. Oliveira, D. Zimdars, Semicond. Sci. Technol. 20 (2005) S266.

B. Ferguson, S. Wang, D. Gray, D. Abbott, X.C. Zhang, Microelectronics Journal 33 (2002) 1043.

B.M. Fischer, M. Walther, P.U. Jepsen, Phys. Med. Biol. 47 (2002) 3807.

B. Fischer, M. Hoffmann, H. Helm, G. Modjesch, P.U. Jepsen, Semicond. Sci. Technol. 20 (2005) S246.

M.J. Fitch, R. Osiander, John Hopkins Apl Technical Digest 25 (2004) 348.

A.J. Fitzgerald, B.E. Cole, P.F. Taday, Journal of Pharmaceutical Sciences 94 (2005) 177.

M. Ge, H. Zhao, T. Ji, X. Yu, W. Wang, W. Li, W. Science in China: Series B Chemistry 49 (2006) 204.

P.A. Gorry, Anal. Chem. 62 (1990) 570.

P.R. Griffiths, J.A. de Haseth, Fourier-transform infrared spectroscopy (New York, John Wiley & Sons, 1986).

I.M. Grigoriev, A.V. Domanskaya, A.V. Podzorov, M.V. Tonkov, Molecular Physics 102 (2004) 1851.

P.Y. Han, M. Tani, M. Usami, S. Kono, R. Kersting, X.C. Zhang, J. Appl. Phys. 89 (2001) 2357.

M. Hangyo, M. Tani, T. Nagashima, International Journal of Infrared and Millimeter Waves 26 (2005) 1661.

M. He, A.K. Azad, S. Ye, W. Zhang, Optics Communications 259 (2006) 389.

L. Ho, R. Müller, K.C. Gordon, P. Kleinebudde, M. Pepper, T. Rades, Y. Shen, P.F. Taday, J.A. Zeitler, Journal of Controlled Release 127 (2008) 79.

L. Ho, R. Müller, K.C. Gordon, P. Kleinebudde, M. Pepper, T. Rades, Y. Shen, P.F. Taday, J.A. Zeitler, European Journal of Pharmaceutics and Biopharmaceutics 71 (2009) 117.

J. Hooper, E. Mitchell, C. Konek, J. Wilkinson, Chem. Phys. Lett. 467 (2009) 309.

Y. Hu, P. Huang, L. Guo, X. Wang, C. Zhang, Physics Letters A 359 (2006) 728.

A.J. Huber, F. Keilmann, J. Wittborn, J. Aizpurua, R. Hillenbrand, Nano Lett. 8 (2008) 3766.

P.U. Jepsen, Optics Express 15 (2007) 14717.

P.U. Jepsen, S.J. Clark, Chem. Phys. Lett. 442 (2007) 275.

M.F. Kimmit, Journal of biological physics 29 (2003) 77.

T. Kleine-Ostmann, K. Pierz, G. Hein, P. Dawson, M. Koch, Electron. Lett. 40 (2004) 124.

M. Knott, New Scientist 2192 (1999) 22.

M. Koch, in: R.E. Miles, X.C. Zhang, H. Eisele, A. Krotkus (Eds.), Terahertz Frequency Detection and Identification of Materials and Objects, second ed., Springer, Netherlands, 2007, p. 325.

M.R. Kutteruf, C.M. Brown, L.K. Iwaki, M.B. Campbell, T.M. Korter, E.J. Heilweil, Chem. Phys. Lett. 375 (2003) 337.

S.P. Langley, Nature (1881).

P. Lena, A.R. King, *Observational Astrophysics*, Springer Ed. (1988)

Y. Li Hor, H.C. Lim, J.F Federici, E. Moore, J.W. Bozzelli, Chemical Physics 353 (2008) 185.

G.F. Liu, X.J. Ma, S.H. Ma, H.W. Zhao, M.W. Ma, M. Ge, W.F. Wang, Chinese Journal of

Chemistry 26 (2008) 1257.

A. Lorber, K. Faber and B. R. Kowalski, Anal. Chem. 69 (1997) 1620.

F.J. Low, W.H. Tucker, Phys. Rev. Lett. 21 (1968) 1538.

H.H. Madden, Anal. Chem. 50 (1978) 1383.

V. Malaterre, M. Perdesen, J. Ogorka, R. Gurny, N. Loggia, P.F. Taday, Eur. J. Pharm. Biopharm. 74 (2010) 21.

A.G. Markelz, A. Roitberg, E.J. Heilweil, Chem. Phys. Lett. 320 (2000) 42.

H. Martens, T. Naes, *Multivariate Calibration* (John Wiley & Sons Ltd, Chichester, 1989).

L. Maurer, H. Leuenberger, International Journal of Pharmaceutics 370 (2009) 8.

C.M. McGoverin, T. Rades, K.C. Gordon, Journal of Pharmaceutical Sciences 97 (2008) 4598.

G. Mie, speziell kolloidaler Metallösungen. Ann. Phys. Leipzig 25 (1908) 377.

D.M. Mittleman, R.H. Jacobsen, M.C. Nuss, IEEE Journal of Selected Topics in Quantum Electronics 2 (1996) 679.

D.M. Mittleman, R.H. Jacobsen, R. Neelamani, R.G. Baraniuk, M.C. Nuss, Appl. Phys. B 67 (1998) 379.

M. Nagel, P.H. Bolivar, M. Bruscherseifer, H. Kurz, A. Bosserhoff, R. Büttner, Appl. Phys. Lett. 80 (2002) 154.

NATO Security through Science Series, Terahertz Frequency Detection and Identification of Materials and Objects, second ed., Springer, Netherlands, 2007.

J.I. Nishizawa, T. Sasaki, T. Tanno, Journal of Physics and Chemistry of Solids 69 (2008) 693.

S.J. Oh, J. Kang, I. Maeng, J.S. Suh, Y.M. Huh, S. Haam, J.H. Son, Optics Express 15 (2009) 3469.

P. Planken P, Nature 456 (2008) 454.

D.F. Plusquellic, K. Siegrist, E.J. Heilweil, O. Esenturk, Chem. Phys. Chem. 8 (2007) 2412.

J. M. Roger, F. Chauchard, V. Bellon-Maurel, Chemometrics and intelligent laboratory systems 66 (2003) 191.

J.M. Roger, B. Palagos, D. Bertrand, E. Fernandez-Ahumada, Chemometrics and Intelligent Laboratory Systems 106 (2011) 216.

M. Rowan-Robinson, Astronomy and Geophysics 47 (2007) 4.31.

H. Rubens, Annalen der Physik 309 (1901) 649.

A. Savitzky, J.E. Golay, Anal. Chem. 36 (1964) 1627.

Y.C. Shen, P.C. Upadhya, E.H. Linfield, A.G. Davies, Vibrational Spectroscopy 35 (2004) 111.

S.M. Short, R.P. Cogdill, C.A. Anderson, AAPS PharmSciTech 8 (2007) Article 96.

C. Sirtori, P. Kruck, S. Barbieri, P. Collot, J. Nagle, M. Beck, J. Faist, U. Oesterle, Appl. Phys. Lett. 73 (1998) 3486.

P.R. Smith, D.H. Auston, M.C. Nuss, IEEE Journal of Quantum Electronics 24 (1988) 255.

S.W. Smye, J.M. Chamberlain, A.J. Fitzgerald, E. Berry, E. Phys. Med. Biol. 46 (2001) R101.

C.J. Strachan, P.F. Taday, D.A. Newham, K.C. Gordon, J.A. Zeitler, M. Pepper, T. Rades, Journal of Pharmaceutical Sciences 94 (2005) 837.

P.F. Taday, R. Soc. Lond. A 362 (2004) 351.

R. Ulrich, Infrared Phys. 7 (1967) 37.

V.P. Wallace, A.J. Fitzgerald, S. Shankar, N. Flanagan, R. Pye, J. Cluff, D.D. Arnone, British Journal of Dermatology 151 (2004) 424.

M. Walther, B.M. Fischer, P.U. Jepsen, PU. Chem. Phys. 288 (2003) 261.

Y. Watanabe, K. Kawase, T. Ikari, H. Ito, Y. Ishikawa, H. Minamide, Optics Communications 234 (2004) 125.

G. Winnewisser, C. Kramer, Space Science Reviews 90 (1999) 181.

S. Wold, Chemometrics and Intelligent Laboratory Systems 30 (1995) 109.

R.M. Woodward, V.P. Wallace, D.D. Arnone, E.H. Linfield, M. Pepper, Journal of Biological Physics 29 (2003) 257.

H. Wu, E.J. Heilweil, A.S. Hussain, M.A. Khan, Journal of Pharmaceutical Sciences 97 (2008) 970.

L. Yang, H. Sun, S. Weng, K. Zhao, L. Zhang, G. Zhao, Y. Wang, Y. Xu, X. Lu, C. Zhang, J. Wu, C. Jia'er, Spectrochimica Acta Part A 69 (2008) 160.

Y. Zhang, X.H. Peng, Y. Chen, J. Chen, A. Curioni, W. Andreoni, S.K. Nayak, X.C. Zhang, Chem. Phys. Lett. 452 (2008) 59.